Oral microbiome characterization of diverse human populations from Africa

Vítor Emanuel da Cunha Araújo

Mestrado em Biodiversidade, Genética e Evolução
Departamento de Biologia
2019

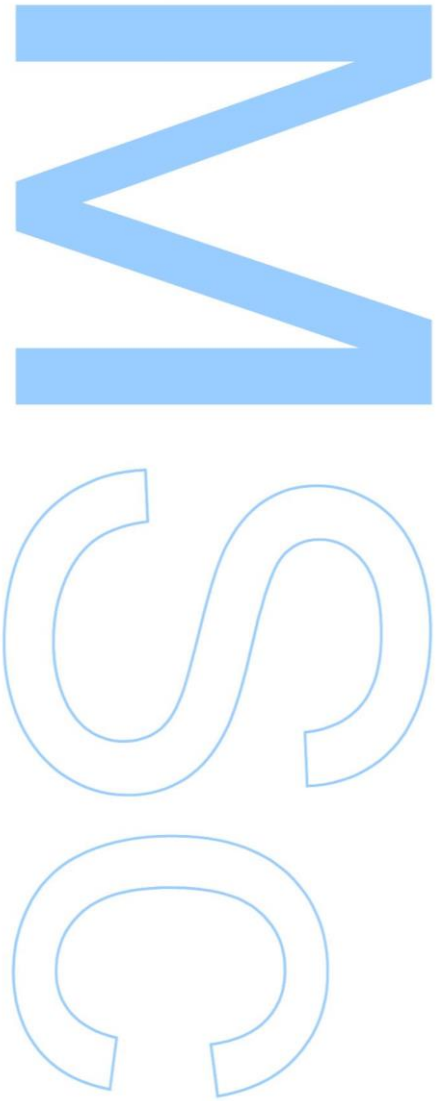
Orientador
Dr. Magdalena Gayà Vidal, Investigador Auxiliar, CIBIO-InBIO

Coorientador
Dr. Jorge Macedo Rocha, Professor Associado, FCUP

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____



Acknowledgments

I would sincerely like to thank my supervisors, without whom this work would not have been possible. Their support and dedication have helped a lot and encouraged me through the hardest periods.

To my supervisor, Magdalena Gayà Vidal, for being always available to help with clear guidance, for her generosity, comprehension, encouragement and overall for the many hours of dedicated work to this thesis.

To my co-supervisor, Jorge Macedo Rocha, for the knowledge that he transmitted me, for the advises and also for the inspiring stories which were always loaded with enthusiasm.

My many thanks to the other members and collaborators of the HUMANEVOL group who have always demonstrated willingness to help in whatever was needed, Anne Maria Fehn, Armando Semo, Sandra Oliveira, Bérenice Alard, João Almeida, Ana Santos, Laia and Beatriz.

I also want to thanks to Mafalda Galhardo for the precious help and time you dedicated when we asked for your advice, and thank you for your sympathy.

To all my colleagues, who shared with me these two years of learning, growing, hard work and joy, a huge thanks for making it unforgettable.

To my lifetime friends, for being understanding and always so supportive despite the different paths each one took in life.

Last but not least, my deepest thank goes to my family. To my Mum and Dad for all the care, for all the sacrifices and for all the comprehension with me. To my sisters and my nephew for the unconditional support and for always being so present.

Resumo

Diferenças no microbioma oral têm sido encontradas entre populações humanas com diferentes dietas, estilos de vida, entre outros fatores. No entanto, a compreensão completa da extensão desta variabilidade está ainda por alcançar e, neste sentido, a caracterização dos microbiomas orais de populações humanas pouco estudadas poderá ser particularmente valiosa.

Aqui, caracterizamos o microbioma oral humano de diferentes populações africanas que vivem em regiões geográficas distintas (São Tomé e Príncipe, Angola, Zimbábue e Moçambique), que têm diferentes idiomas (Khoisan, Bantu e Crioulo) e métodos de subsistência (agro-pastoris, pastoris, peripatéticos e forrageadores) usando a porção de 'reads' não humanos obtidos através de uma abordagem de Sequenciação de Captura do Exoma Expandido a partir de saliva e raspagens da mucosa das bochechas.

No geral, encontramos diferentes padrões de medidas de diversidade, com as populações agro-pastoris a apresentarem menor diversidade intra-individual e maior diversidade inter-individual. Por sua vez, os Sekele, o único grupo de forrageadores, apresentou a maior diversidade intra-individual. Para além disto, as análises de agrupação e ordenação discriminaram a maior parte das populações agro-pastoris de Angola e de São Tomé e Príncipe das restantes populações não agro-pastoris, nas quais, nenhuma diferenciação foi detetada entre os grupos pastoris, peripatéticos e forrageadores angolanos. A diferenciação entre os grupos agro e não agro-pastoris deve-se principalmente ao enriquecimento de diversos táxons neste último, tal como, *Atopobium sp*., *Solobacterium moorei* e *Veillonella dispar*. Propomos que a separação destes grupos pode ser devida a diferentes dietas e/ou estatutos económicos das populações estudadas.

Palavras-chave

Microbioma oral, Saliva, Métodos de subsistência, Captura do Exoma, Populações Africanas.

Abstract

Differences in the oral microbiome have been found between human populations having different diets, lifestyles, among other factors. However, a full comprehension of the extent of this variation has not yet been accomplished, and the characterization of the oral microbiomes from understudied human populations might be particularly valuable in this regard.

Here, we successfully characterized the human oral microbiome of different African populations living in distinct geographic regions (São Tomé and Príncipe, Angola, Zimbabwe and Mozambique), having different languages (Khoisan, Bantu and Creole) and subsistence methods (agropastoralists, pastoralists, peripatetic and foragers) by using the portion of non-human reads obtained through an Expanded Exome Capture Sequencing approach from saliva and cheek scraps.

Overall, we found different patterns of diversity measures, showing most agropastoral populations lower intra-individual and higher inter-individual diversity compared to the remaining groups. The Sekele, the only foragers, had the highest intra-individual diversity. Furthermore, the cluster and ordination analyses discriminated most Angolan and São Tomé and Príncipe agropastoral populations from the remaining non-agropastoral populations where no differentiation was detected between the Angolan pastoralists, peripatetics and foragers. The differentiation between the agro and non-agropastoral clusters are mainly due to enrichment in the latter of several taxa such as, *Atopobium sp*., *Solobacterium moorei* and *Veillonella dispar.* We propose that the separation of these two clusters could be due to different diets and/or economic-status of the populations studied.

Keywords

Oral microbiome, Saliva, Subsistence methods, Exome capture, African populations.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DNA - Deoxyribonucleic Acid

rDNA – Ribosomal Deoxyribonucleic Acid

RA - Rheumatoid Arthritis

FASTQ – Text-based file that contains sequence alignment data and correspondent quality scores

BWA - Burrows-Wheeler Aligner

BAM - Binary Aligned/Mapped

HMP - Human Microbiome Project

NCBI - National Center for Biotechnology Information

BLAST - Basic Local Alignment Search Tool

PCA - Principal Component Analysis

NMDS - Non-Metric Multidimensional Scaling

Pheatmap - Pretty Heatmap

PERMANOVA - Permutational Multivariate Analysis of Variance

Anova - Analysis of Variance

1. Introduction

1.1. Oral microbiome

The number of human cells that compose the human body is of the same order of magnitude as the number of microbial cells that colonize it (Sender et al., 2016). The microbial cells occupy different habitats of the human body. One of the largest and most complex of these habitats is the oral cavity (Wade, 2013; Chen and Jiang, 2014). The group of microorganisms living in the oral cavity are designated by oral microbiota and the set of their genomes the oral microbiome (Chen and Jiang, 2014).

The oral microbiota is heterogeneously distributed in different sites of the oral cavity (Mager et al., 2003), nevertheless, the microorganisms present in saliva are partially shared with those found in other sites such as tongue dorsum, buccal mucosa and plaque (Mager et al., 2003; Eren et al., 2014). As a consequence, saliva has been widely used to study the human oral microbiome (Takeshita et al., 2014; Grassl et al., 2016; Nakano et al., 2018).

1.2. Sequencing approaches

To study the microbiome, there are two next-generation sequencing (NGS) approaches commonly used: i) amplicon sequencing (amplification and subsequent sequencing of a target DNA fragment) of the 16S rDNA gene (gene present in all bacteria), and ii) shotgun sequencing (random sequencing of the DNA from a sample). While the amplicon sequencing is a most cost-effective approach, the shotgun sequencing has the advantage of being able to detect other microorganisms than bacteria, such as fungi and viruses. Furthermore, shotgun sequencing studies reportedly detect more diversity than amplicon ones (Poretsky et al., 2014; Ranjan et al., 2017). Interestingly, Kidd et al., (2014) characterized the oral microbiome from saliva samples by focusing on reads that did not align to the human genome obtained with a Human Exome Capture Sequencing, an approach that target the human exome. The authors reported that the structure and abundance of microorganisms obtained seem to be consistent with that obtained through traditional shotgun metagenomic sequencing (Kidd et al., 2014).

1.3. Oral microbiome across different populations

Studies on oral microbiome have revealed a great diversity of microbial communities, both in terms of abundance distribution and composition. These differences have been related to diseases as well as to different environments, diets and other factors.

On one hand, the oral microbiome has been related to a multitude of oral and systemic diseases, including caries, periodontitis, diabetes, obesity, liver diseases, colon cancer, oral and pancreatic cancer and RA (a systematic autoimmune disease) (Lu et al., 2019). The incidence of some of these diseases varies among human groups, which may result from intrinsic differences in their oral microbiota (Gupta et al., 2017).

On the other hand, the characterization of the oral microbiota in a variety of human populations revealed differences in the oral microbiome profile related with distinct diets, lifestyles, and environmental conditions. Nasidze et al., (2009) compared 12 worldwide locations and found the biggest differences between individuals' oral microbiomes in Congo. By contrast, the individuals from Georgia and Turkey presented the most similar oral microbiomes. In relation to individuals of other countries, individuals from Congo showed an increased frequency of several genera being the most striking case the genus *Enterobacter* (in Congo corresponded to 28% of the sequences while in California, China, Germany, Poland, and Turkey was absent). Overall, the authors found that the oral microbiome was not strongly influenced by geography. Kidd et al., (2014) found that Khoisan populations possess higher oral pathogenic microbial load than the one observed in North Americans from the Human Microbiome Project (Methé, 2012). These authors hypothesised that the observed differences could result from "limited access to dental care, antibiotics and/or absence of water fluoridation among the KhoeSan" (Kidd et al., 2014).

Several studies found that significantly different oral microbiomes were shaped by the host genetics. Mason et al., (2013), studied individuals from the four major ethnicities of the United States, and showed that microbial communities from saliva and subgingival biofilms were characteristic from the individuals' ethnicity, the degree of association was enough to a machine-learning classifier being able to discriminate individuals between ethnicities based on their microbial profiles. It was suggested that more similar oral microbiomes between individuals of the same ethnic group could result from more similar tooth and root morphologies and innate immune responses to infectious agents. In line with this study, Blekhman et al. (2015) found a significant association between the human genetic variations and the microbial composition of several sites of the oral cavity by using data from the Human Microbiome Project. Demmitt et al., (2017) showed that, independently of cohabitation status during a sampled period of 2-7 years, twins maintained a relatively more similar oral microbiome than unrelated people. These authors found two human loci on chromosomes 7 and 12, which influence the microbial phenotypes.

Other studies attribute most of the differences of the oral microbiome to the host lifestyle and particularly to their diet. Nasidze et al., (2011) found that Batwa Pygmies, a former hunter-gatherer group from Uganda, possess a significant higher microbial diversity than the agricultural groups from Sierra Leone and from the Democratic Republic of Congo, and suggested that the diet (protein-rich in Batwa Pygmies) was the main driver of the observed differences. These authors also report that Batwa Pygmies populations possess 40 microbial genera which have never been described in the Human oral cavity, reinforcing the necessity to analyse the oral microbiome in more detail, and in more diverse human populations (Nasidze et al., 2011). Li et al., (2014) demonstrated significant differences in the oral microbiome diversity of German and African populations. The oral microbiome of Germans was very diverse within individuals but quite similar between them, whereas in African populations the pattern was the opposite. These authors suggest the variability of the diet and the degree of human concentration as the main drivers of the observed pattern (Li et al., 2014). Takeshita et al., (2014) compared the oral microbiome of genetically similar human populations from South Korea and Japan. The Japanese oral microbiome was characterized by "higher proportions of *Prevotella* and *Veillonella* and lower proportions of *Neisseria* and *Haemophilus*" comparatively to the oral microbiome of the Koreans, and these differences were thought to be correlated with the worst periodontal status of the Japanese. The authors suggest that the diet, spicier and saltier in the Koreans, was the main driver of the observed differences in the oral microbiome. Lassalle et al. (2018) studied the oral microbiome of hunter-gatherer and farmer populations from the Philippines and compared them with that of individuals from a western lifestyle. The authors found that the oral microbiomes were significantly correlated with the subsistence strategy. While Hunter-gatherers were enriched in *Neisseria* species, westerns were enriched in *Haemophilus,* and farmers fell in between this gradient. These results confirm that major shifts in the human oral microbiome composition occurred in line with marked dietary shifts, being the first one, with the advent of agriculture when the consumption of food with high levels of carbohydrates increased, and the second, with the Industrial Revolution and the advent of industrially processed flour and sugar (Adler et al., 2013; Schnorr et al., 2016; Gupta et al., 2017).

These studies have shown that the oral microbiome is highly diverse; in particular, the oral microbiome of groups that maintain traditional lifestyles seems to be widely different from the well-studied oral microbiome of individuals from westerns civilization. So, to gain knowledge about the oral microbiome, more studies need to be done of populations having different subsistence modes and from understudied world regions.

In this context, we analyse the oral microbiome of 95 individuals from several populations living in different areas of Africa and with a high cultural and biological diversity. This work will allow us to contribute valuable oral microbiome data from an understudied region, and to investigate the influence of different factors such as environment and lifestyle to obtain a more accurate picture of the general composition of the human oral microbiome.

1.4. Objectives

The present study aims to contribute to the general knowledge of the oral microbiota by analysing the human oral microbiome composition of several populations from Africa living in different geographic areas, with different subsistent modes and population histories. In this work, we take advantage of the sequence data available from 95 individuals from different African populations obtained through an Expanded Exome Capture Sequencing approach, as in Kidd et al., (2014).

In order to achieve the main goal of this study, four more specific objectives were defined:

I) Characterize the oral microbiome of each individual by identifying the species and genera and their abundance distribution.

II) Calculate the diversity of the oral microbiome both within and between individuals.

III) Compare the oral microbiome profiles and diversity levels of the different populations. Identify which species and genera are the most differentiated between groups.

IV) Investigate whether differences on the oral microbiome exist depending on the characteristics of the studied populations (lifestyles, environment, genetic background).

2. Material and Methods

2.1. Samples

Saliva or cheek scrap samples along with ethnographic data referent to language and genealogical aspects were collected with written informed consent from a total of 96 African individuals. The sampled individuals belong to 16 populations from 4 African countries: Angola (10 populations), Mozambique (1 population), São Tomé and Príncipe (3 populations) and Zimbabwe (2 populations), which location is shown in Figure 1.

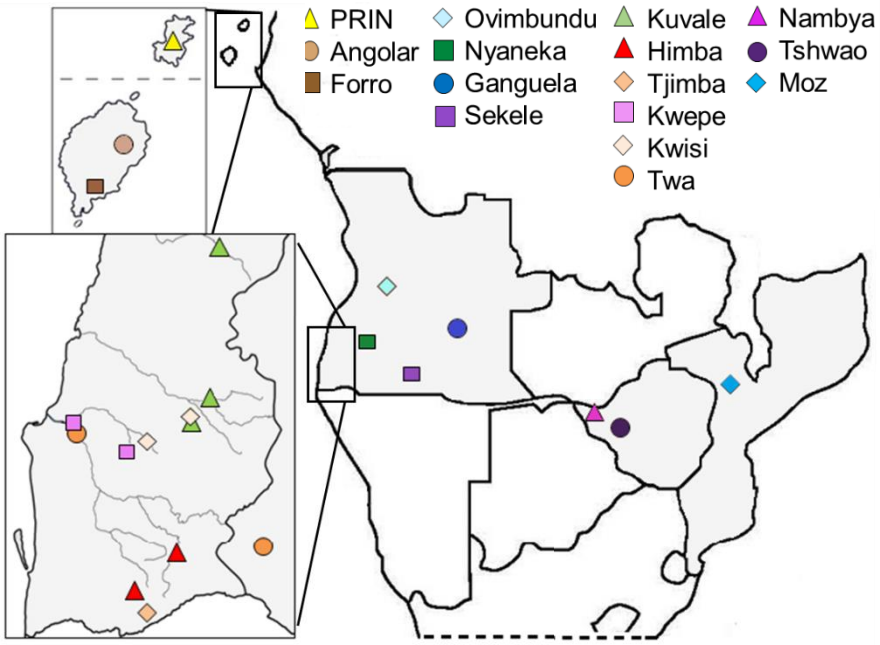


Figure 1. Geographic representation of the location of the studied populations in (A) São Tomé and Príncipe, (B) Angola, (C) Zimbabwe and (D) Mozambique.

In addition to living in distinct geographic regions, the sampled populations present several other distinct characteristics, namely, different population histories, languages and subsistence patterns. Details about these populations can be consulted in Table 1.

Table 1. Country, language, subsistence pattern, collected sample type and number of individuals for each population analysed.

Population	Country	Language family	Subsistence pattern [1]	Sample type	Number of individuals
Angolar	São Tomé and Príncipe	Creole	Agropastoralists	Cheek scraps	9
Forro	São Tomé and Príncipe	Creole	Agropastoralists	Cheek scraps	8
Príncipe	São Tomé and Príncipe	Creole	Agropastoralists	Cheek scraps	8
Ovimbundu	Angola	Bantu	Agropastoralists	Saliva	5
Nyaneka	Angola	Bantu	Agropastoralists	Saliva	5
Ganguela	Angola	Bantu	Agropastoralists	Saliva	5
Kuvale	Angola	Bantu	Pastoralists	Saliva	7
Kwepe	Angola	Khoe-Kwadi (Khoisan)/Bantu [2]	Pastoralists	Saliva	7
Kwisi	Angola	Bantu	Peripatetic	Saliva	8
Twa	Angola	Bantu	Peripatetic	Saliva	7
Himba	Angola	Bantu	Pastoralists	Saliva	7
Tjimba	Angola	Bantu	Peripatetic	Saliva	5
Sekele	Angola	Kx'a (Khoisan)	Foragers	Saliva	7
Tshwao	Zimbabwe	Khoe-Kwadi (Khoisan)	Peripatetic [3]	Saliva	5
Nambya	Zimbabwe	Bantu	Agropastoralists	Saliva	1
Mozambique	Mozambique	Bantu	Agropastoralists	Cheek scraps	2

[1] The term peripatetic aims to classify the low-status, primarily non-food-producing populations Kwisi, Twa, Tjimba and Tshwao.

[2] Khoe-Kwadi was the former language of the Kwepe, but nowadays this group speaks Bantu.

[3] Tshwao lived as foragers but nowadays are better classified as peripatetics.

2.2. DNA extraction and sequencing

DNA extraction was conducted for saliva and cheek scraps as previously described (Quinque et al., 2006). Library preparation and expanded exome enrichment were performed using Nextera® Rapid Capture Enrichment kit by Illumina.

The 96 individuals were sequenced in three different sequencing runs. One run sequenced 24 indexed samples corresponding to individuals of São Tomé and Príncipe in two lanes using Illumina's HiSeq 1500 System in Rapid Run mode. Other 24 samples from Angolan populations were also sequenced using two lanes in an Illumina's HiSeq 1500 System in Rapid Run mode. The other 48 indexed samples were sequenced in four lanes, using Illumina's HiSeq 1500 System in High Output Run mode.

The resulting FASTQ files were processed in order to remove low-quality reads by applying a filter for Phred Quality Score of 30 (Q30) with Sickle (v1.33) (Joshi and Fass, 2011) in pair-end mode. Reads that passed the quality filter were aligned to the human genome hg19 using the *-mem* option of Burrows-Wheeler Aligner (BWA) software (v0.7.15) (Li, 2013). After the alignment one of the Tjimba samples presented very low

number of reads and it was excluded from this work. The BAM files resulting from the alignment contained all sequenced reads, those that aligned to the human genome and those that did not align. We used these BAM files to proceed to the metagenomic analysis.

2.3. Metagenomic pipeline

As a first step of the metagenomic pipeline, from the BAM files containing all information about the alignment of the reads to the human genome hg19, we extracted the non-human reads, that is, the unmapped reads, using the option *view -b -f 4* of the software SAMtools (Li et al., 2009).

The non-human reads were further subjected to a quality filtering suitable for the metagenomic analysis. We used PRINSEQ tool (Schmieder and Edwards, 2011) to remove reads with less than 50 bp, reads with a mean quality score less than 25 and reads which were exact duplicates in accordance with the criteria used in Kidd et al., (2014). Since PRINSEQ works with FASTQ files, the BAM files were first converted using BEDtools (Quinlan and Hall, 2010).

In order to identify and characterize the non-human organisms present in the samples, we used the fragment recruitment approach, which consists in mapping metagenomic reads against a set of selected references genomes (Rusch et al., 2007). Thus, we downloaded the microbiome reference genomes from the Human Microbiome Project (HMP) (NCBI BioProject PRJNA28331) (November 19, 2018) and created a BLAST database using the option *makeblastdb* of the software BLAST+ (Altschul et al., 1990). The high-quality non-human reads were blasted against this database using the option *blastn* of the software BLAST+ and the best hit for each read was retained. In accordance with the work of Kidd et al., (2014), for the species-level binning (i.e. to consider that a read was in fact amplified from a specific species) we required that the alignment covered at least 75% of the read length, and the sequences were at least 95% identical. We also applied other less stringent criteria for read recruitment as in Kidd et al., (2014) in order to compare the relative abundance of reads obtained at each restriction level.

2.4. Data analysis

The abundance of each microbial species in each individual was inferred according to the number of metagenomic reads aligned against each species and a species abundance table was constructed. Species that did not recruit any read in any sample were removed. A genera abundance table was also constructed from the species

abundance table by merging species of the same genus, and both tables were used for the subsequent analyses. The correlation between the number of taxa (species and genera) detected in each individual and the correspondent number of reads was calculated using a Spearman rank correlation test using the function “cor.test” from the package stats v3.6.1 (R Core Team, 2018).

Two diversity measurements were estimated from our data, the alpha (diversity within individuals) and the beta (diversity between individuals) diversities. Prior to alpha and beta diversities calculations, we performed a square-root transformation of our data in order to attenuate the influence of highly-abundant species. This transformation is usually applied to count variables (Osborne, 2002). Alpha diversity was calculated using the Shannon index (Shannon, 1948) with the function “diversity” from the package vegan v2.4-2 (Oksanen et al., 2019). Beta diversity was calculated using the Bray–Curtis dissimilarity (Bray and Curtis, 1957) with the function “vegdist” also from the package vegan v2.4-2.

To explore how the individuals analysed clustered according to their microbiome profiles, two ordination methods were applied, principal component analysis (PCA) and non-metric multidimensional scaling (NMDS). For the PCA analysis, we first carried out a variance-stabilizing transformation of our data based on the negative binomial model using DESeq2 (Love et al., 2014; McMurdie and Holmes, 2014). This transformation intends to reduce or eliminate the dependence of the variance on the mean so that all variables have the same variance. Here, each variable corresponds to the number of reads recruited for each microbial taxon in the different individuals. PCA was performed using the function “dudi.pca” from the package ade4 v1.7-13 (Dray and Dufour, 2007). To visualize the PCA outputs we used the function “fviz_pca” from the package factoextra v1.0.5 (Kassambara and Mundt, 2017). NMDS was performed using the function “metaMDS” from the package vegan v2.4. Due to the large discrepancy of the number of recruited reads for a determined microbial taxon, a square root-transformation followed by a Wisconsin double standardization was automatically applied by metaMDS, and then the Bray-Curtis dissimilarity matrix was calculated. The functions “ordiplot”, “ordihull” and “orditorp” from the package vegan v2.4-2 were used to display the NMDS outputs.

To visualize whether differences in the distribution of species/genera existed among individuals, we created heatmaps of species and genera abundance and conducted the hierarchical clustering of individuals. For this analysis we used the Bray–Curtis dissimilarity values from the data transformed with DESeq2. Heatmaps were created using the function “pheatmap” from the package pheatmap v1.0.12 (Kolde, 2019),

and the hierarchical clustering of individuals based on the Bray-Curtis dissimilarity was calculated with the clustering method ward.D2 (Murtagh and Legendre, 2014).

PCA, NMDS, heatmap and the associated hierarchical cluster analysis, were performed for 3 sets of individuals with respect to a geographical criterion: one dataset considering all the 95 individuals analysed, a second dataset considering only individuals from southern Africa (Angola, Zimbabwe and Mozambique), and the third one only considering individuals from Angola.

In order to investigate whether the oral microbiome profiles of the studied populations were statistically different according to their characteristics (Table 1), a Permutational multivariate analysis of variance (PERMANOVA; Anderson et al., 2001) was conducted using the function “adonis” from the package vegan v2.4-2. Significance was tested using 9,999 permutations. The PERMANOVA analysis was based on the Bray–Curtis dissimilarity, which was calculated from the data transformed with DESeq2. We grouped populations according to different criteria and compared populations i) with different type of collected samples, ii) speaking different language families, ii) from different countries, and iii) with different subsistence patterns. Regarding the subsistence patterns, taking into account that the Tshwao are former foragers but nowadays they live as agropastoralists, we carried out two different analyses including them into the agropastoral or the forager group. PERMANOVA compares groups assuming in the null hypothesis that the centroids and the dispersion of the groups are equivalent. Thus, when the null hypothesis of PERMANOVA is rejected we cannot discriminate what differ between groups, if the centroids, the dispersions or both. For comparisons in which the null hypothesis of the PERMANOVA was rejected, we calculated the dispersions of the groups using the function "betadisper" from the package vegan v2.4-2 and tested its significance using the function "Anova" from the package car (Fox and Weisberg, 2019).

All the analyses and graphs performed in this study were conducted in R studio version 3.5.2 (R studio team, 2016).

3. Results

3.1 Oral microbiome quantifications

From the total number of reads obtained for each sample with the Expanded Exome Capture Sequencing approach (≈ 32 million reads per individual), 2.18% (1.64% in cheek scraps and 2.43% in saliva) were unmapped since they did not align to the human

genome hg19 (≈ 700,000 reads per individual). After the exclusion of low-quality reads with Prinseq tool, an average of 630,000 high-quality non-human reads remained for each individual (Figure 2), these reads were aligned against the microbiome reference genomes of the Human Microbiome Project (HMP) (NCBI BioProject PRJNA28331) and three stringency criteria for reads recruitment were applied (Figure 3).

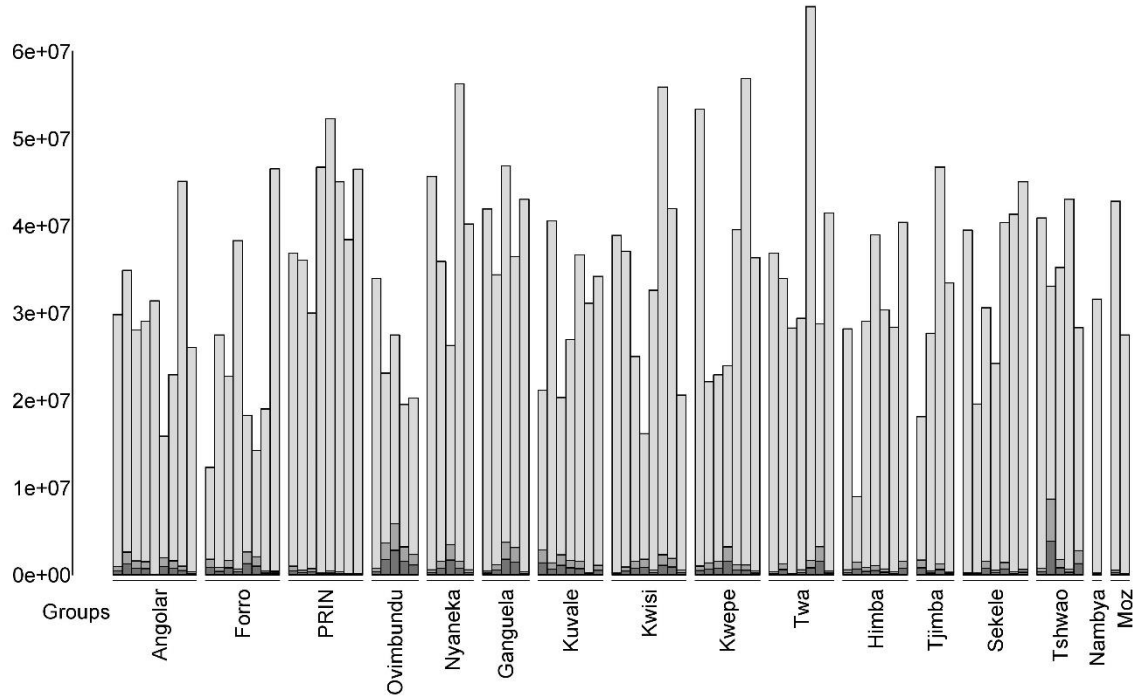


Figure 2 | Number of reads per individual at different stages of the bioinformatic pipeline. Light grey bars represent the total number of reads obtained for each sample after sequencing. Grey bars represent the number of reads unmapped after the alignment with the human genome and dark grey bars represent the number of unmapped reads that meet the quality criterion defined by Prinseq tool.

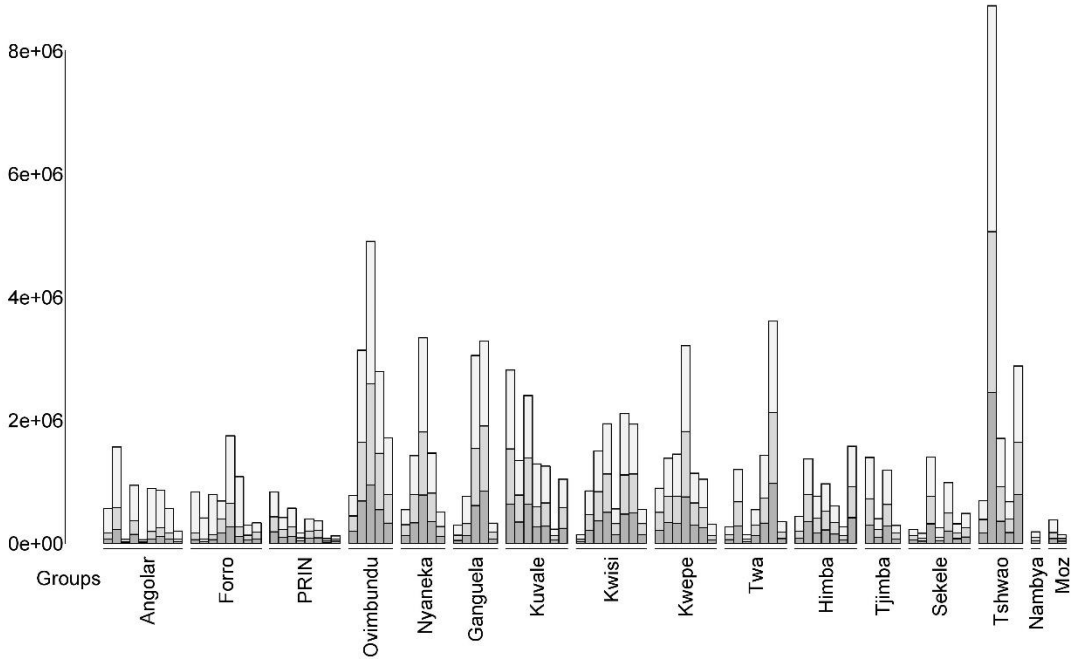


Figure 3 | Number of reads aligned to the microbiome database at different restriction levels. Light grey bars represent reads with at least 50% identity to microbial reference genomes. Grey and dark grey bars represent, respectively, reads with at least 80% and 95% of identity to microbial reference genomes and which aligned in at least 75% of its length.

Focusing on the reads that aligned under the most stringent criteria, in individuals whose DNA was extracted from saliva, we detected an average of 276 species (ranging from 203 - 372) and 105 genera (69 -150). In individuals whose DNA was extracted from cheek scraps, we detected an average of 206 species (ranging from 138 - 274) and 76 genera (49 - 107) (Figure 4 and 5; Figure S1 and S2, Appendix).

Regarding the number of microbial species identified, the groups from São Tomé and Príncipe (Angolar, Forro and Príncipe), Mozambique and the Nambya individual from Zimbabwe, recruited, on average, a considerably lower number of microbial species per individual compared to the other groups (ranging from 194 - 234). However, we need to take into account that only two Mozambican individuals and one Nambya were included in this study. On the other hand, the populations Ovimbundu, Kwepe and Nyaneka are those that recruited, on average, a higher number of microbial species per individual (285 - 300). Regarding the population variance in the number of microbial species identified per individual, the Himba, Kuvale, Mozambique, Príncipe and Tshwao present the lowest values (391 - 730) whereas Ganguela and Ovimbundu present the highest ones (2406 - 2838) (Figure 4).

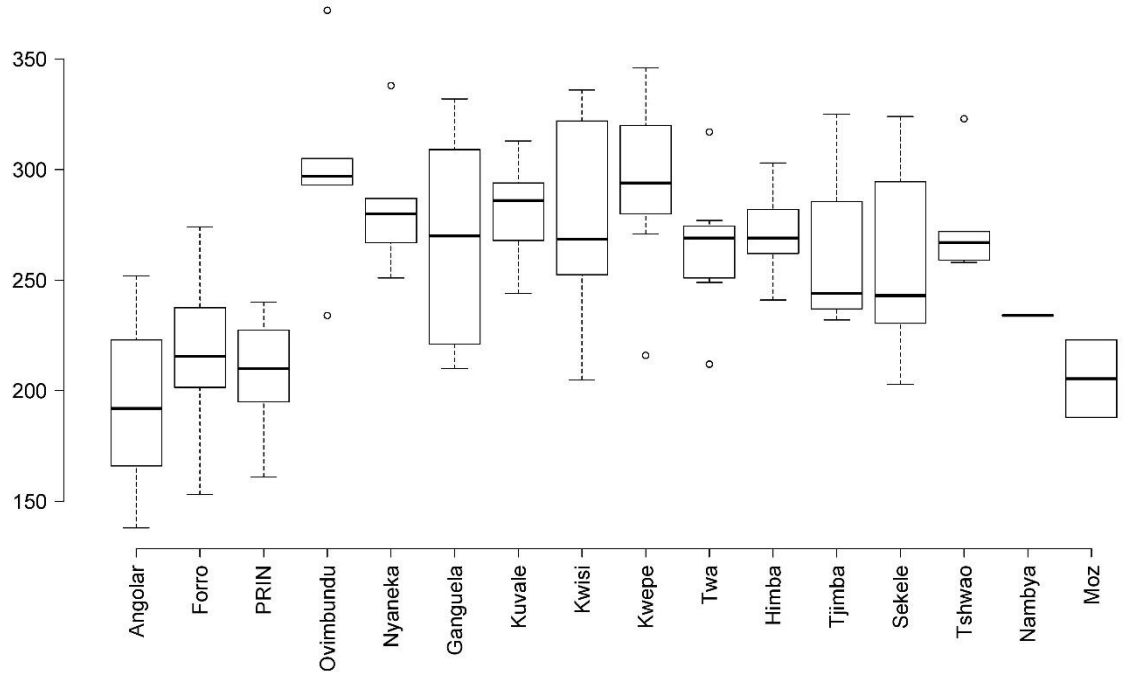


Figure 4 | Box plots showing the variation in the number of microbial species identified in the individuals of each population.

Considering the number of microbial genera identified in each population, we found a similar pattern with that of the species. The populations from São Tomé and Príncipe, Mozambique and the Nambya individual recruited a considerably lower number of microbial genera per individual (ranging from 74 - 81) whereas Ovimbundu, Nyaneka and Kwisi recruited, on average, a higher number (112 - 117). Regarding the population variance in the number of microbial genera identified per individual, a more uniform pattern than that of the species was found. The Himba, Mozambique and Kuvale present the lowest values (38 - 88) whereas the highest variance was found in Sekele (480) (Figure 5).

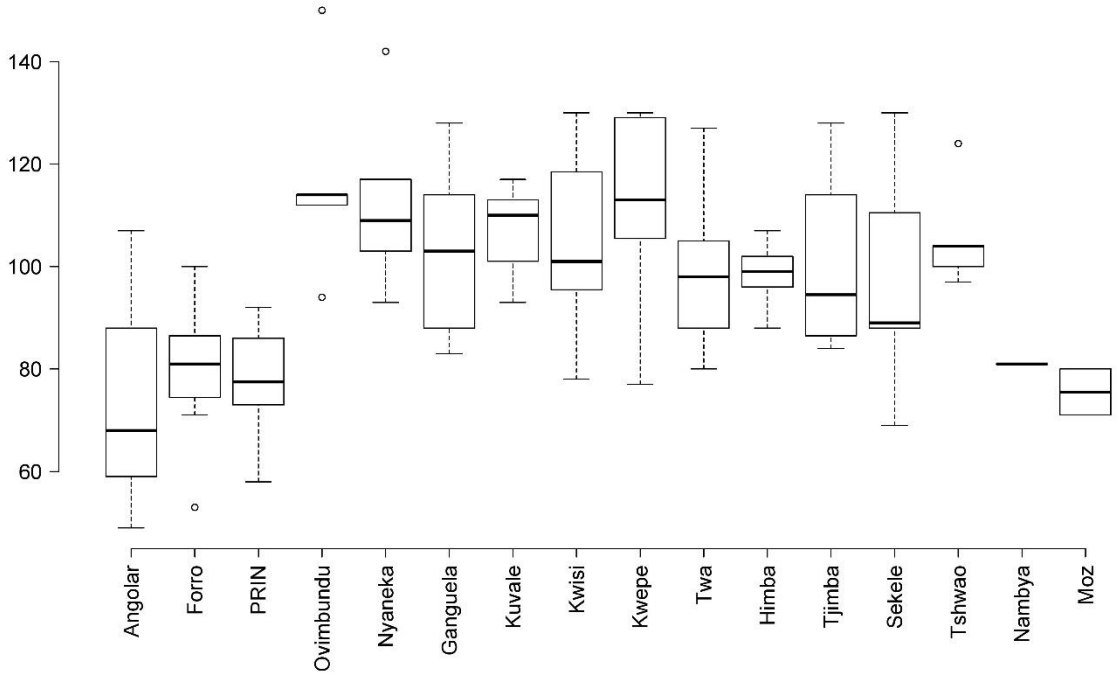


Figure 5 | Box plots showing the variation in the number of microbial genera identified in the individuals of each population.

To check whether the differences in the number of species/genera detected per individual were influenced by different number of sequencing reads, we calculated a Spearman correlation. The results indicated a significant positive association between the number of microbial reads obtained for each individual and the correspondent number of species (r_s = 0.839, p-value < 2.2e-16) and number of genera identified (r_s = 0.851, p-value < 2.2e-16) (for species see Figure 6; for genera see Figure S3, Appendix).

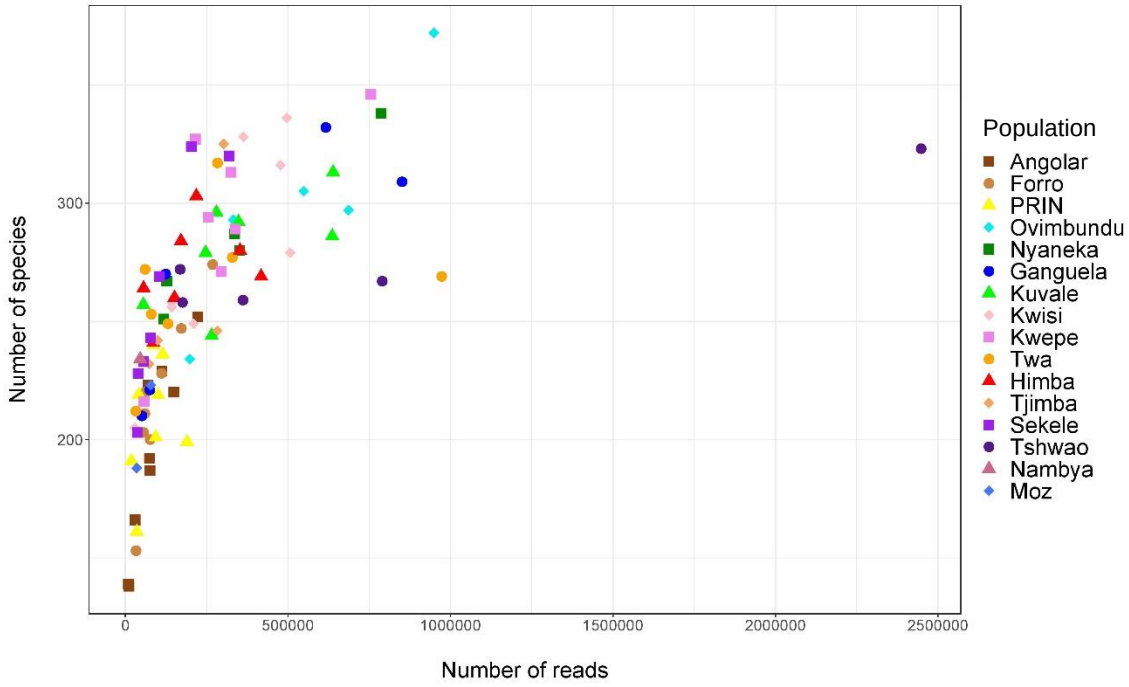


Figure 6 | Relation between the number of reads obtained for each individual and the correspondent number of microbial species identified. Spearman correlation between these two variables: r_s = 0.839 p-values < 2.2e-16.

Regarding the pattern of reads distribution between microbial organisms, we observe that few species recruit most of the reads. After calculating the relative frequencies of each taxa, the 20 most abundant species recruit 68.7% of the reads while the 100 most abundant species recruit 95% of the reads. The 10 most frequent genera are: *Neisseria*, which recruits 21% of the reads, *Streptococcus* (17.3%), *Prevotella* (12%), *Rothia* (7.5%), *Porphyromonas* (6.3%), *Haemophilus* (4.8%), *Actinomyces* (4.3%), *Veillonella* (3%), *Enterobacter* (2.7%) and *Capnocytophaga* (2.4%) (Figure 7).

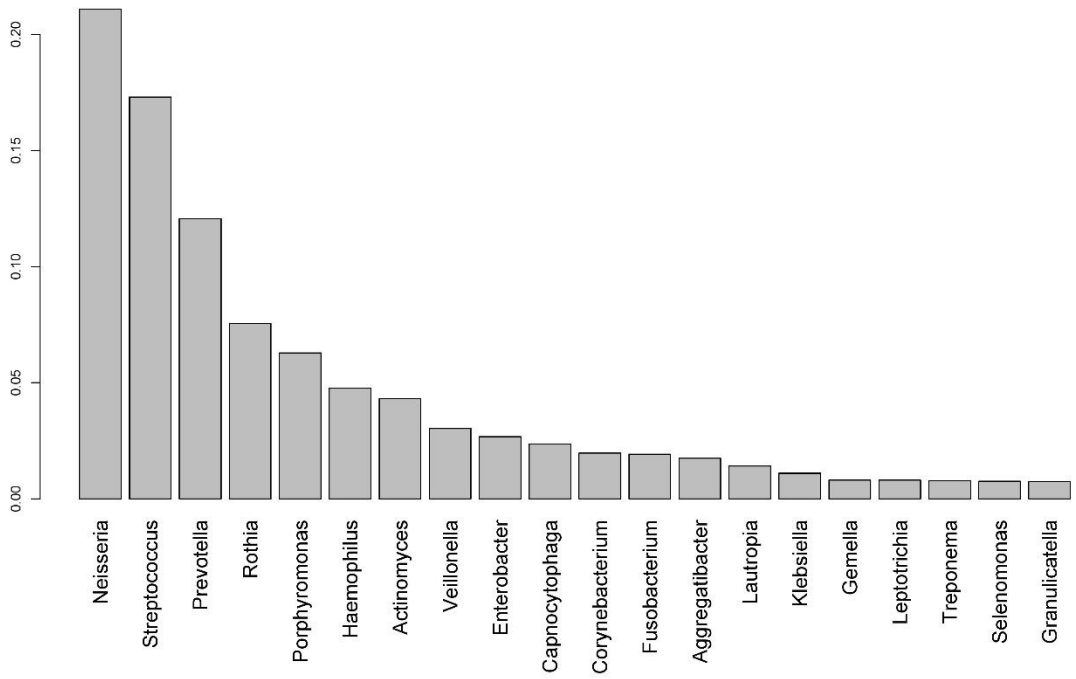


Figure 7 | Relative frequency of the 20 most frequent genera in the populations studied.

3.2 Oral microbiome diversity

Regarding the values of alpha diversity calculated with the Shannon index, which measures how abundant and evenly the microbes are distributed in a sample, at the species level, the populations from São Tomé and Príncipe (Angolar, Forro and Príncipe), Mozambique and the Tshwao from Zimbabwe present the lowest diversity values (mean values between 4.53 and 4.64). In contrast, the Sekele, along with Kwisi, Kwepe, Twa, Himba, Tjimba and Nambya, present the highest values (mean values between 4.8 and 4.88), in these populations the mean and, particularly, the median values are relatively homogenous. The Ovimbundu, Ganguela, Nyaneka and Kuvale present intermediate alpha diversity values (mean values between 4.68 and 4.77) (Figure 8). At the genus level, although a similar pattern is observed, as less taxonomic entities are accounted, the range of values is smaller. On average, Angolar individuals present the lowest values of alpha diversity (3.33) while Sekele present the highest (3.66) (Figure S4, Appendix).

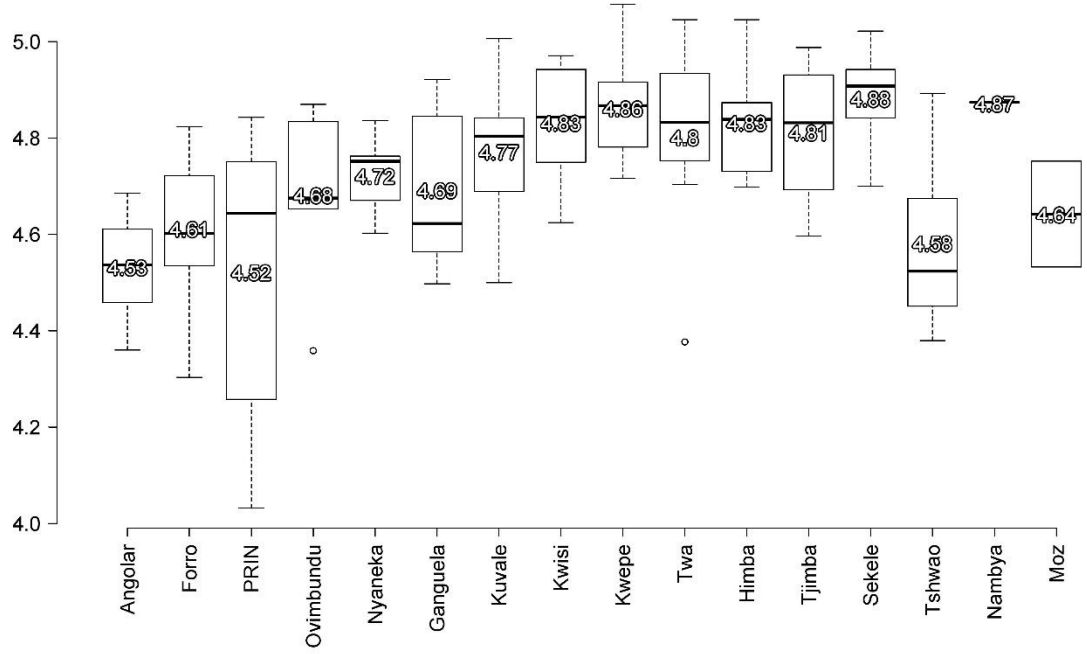


Figure 8 | Box plots showing the distribution of alpha diversity values (Shannon index) calculated for the individuals of each population at the species level. Numbers inside box plots correspond to populations mean values.

Regarding the values of beta diversity calculated as Bray-Curtis dissimilarity, which measures pairwise differences in species composition between individuals and for which a value of 0 means the two sites share all the species and a value of 1 means they do not share any species, we observe that individuals from the populations of São Tomé and Príncipe (Angolar, Forro and Príncipe), along with Ovimbundu, Ganguela and Tshwao present the most differentiated oral microbiomes. On the other hand, individuals from the groups Kuvale and Himba present the most similar ones (Figure 9). At the genus level, the groups Angolar, Ganguela and Tshwao are the ones whose individuals present the most differentiated oral microbiomes while the individuals with more similar ones are, consistently, from the Kuvale and Himba (Figure S5, Appendix).

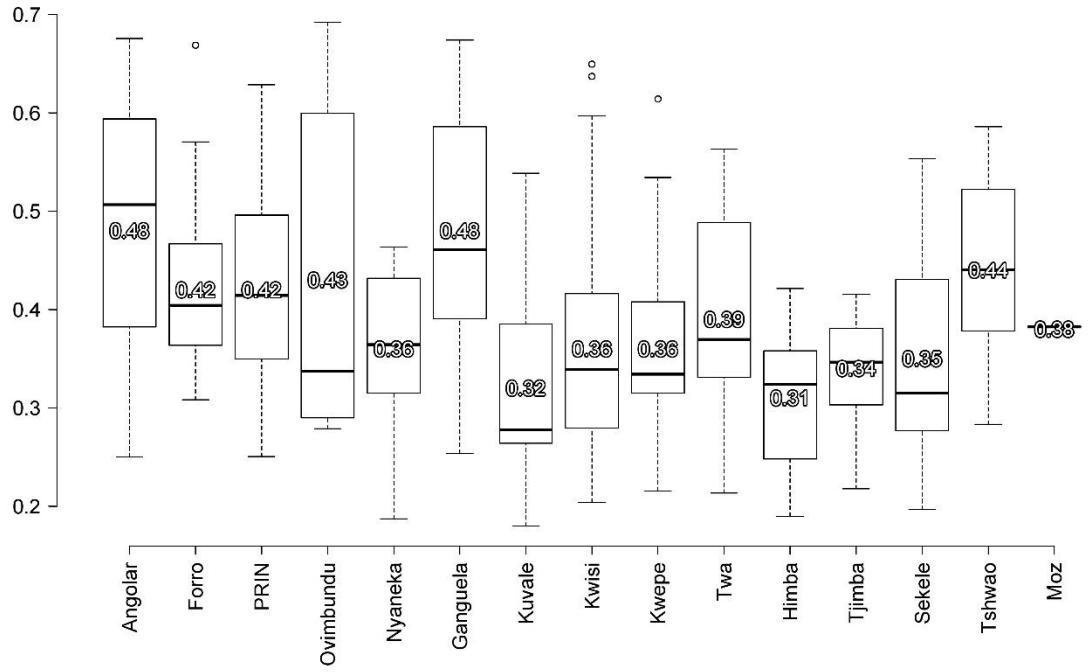


Figure 9 | Box plots showing the distribution of beta diversity values (Bray–Curtis dissimilarity) calculated between pairs of individuals of each population at the species level. Numbers inside box plots correspond to populations mean values.

3.3 Ordination and cluster analysis based on microbial profiles

In order to evaluate how individuals relate to each other based on their microbial profiles we performed a Principal Component Analysis (PCA). In the PCA based on species frequencies (Figure 10), the first three components explain roughly 25% of the variance (PC1 = 12.3%, PC2 = 7.4% and PC3 = 5%). In PC1-PC2 (Figure 10A) we can observe that populations are roughly separated into two groups. On one hand the agropastoral populations from São Tomé and from Angola, and on the other hand, the remaining populations, including the forager Sekele, the pastoralist and peripatetic groups from Angola, and the Tshwao and the Nambya from Zimbabwe. Mozambique and Príncipe appear in a more intermediate position. Considering population midpoints in PC1, the maximum separation is between the Kwepe and Nyaneka. PC2 clearly separates some individuals from the agropastoral Ovimbundu, Angolar and Príncipe populations. In PC3 we observe a further separation of some individuals from Príncipe. Here, the agropastoral populations from Angola are separated from those of São Tomé and Príncipe while Mozambique occupies an intermediate position (Figure 10B). The PCA based on genera frequencies roughly present the same pattern as the species one. The main difference is that in PC3 we observe a marked separation of Tshwao from the other populations. In this PCA the first three components explain roughly 26% of the variance (PCA1 = 13.2%, PCA2 = 7% and PCA3 =5.6%) (Figure 10C and D).

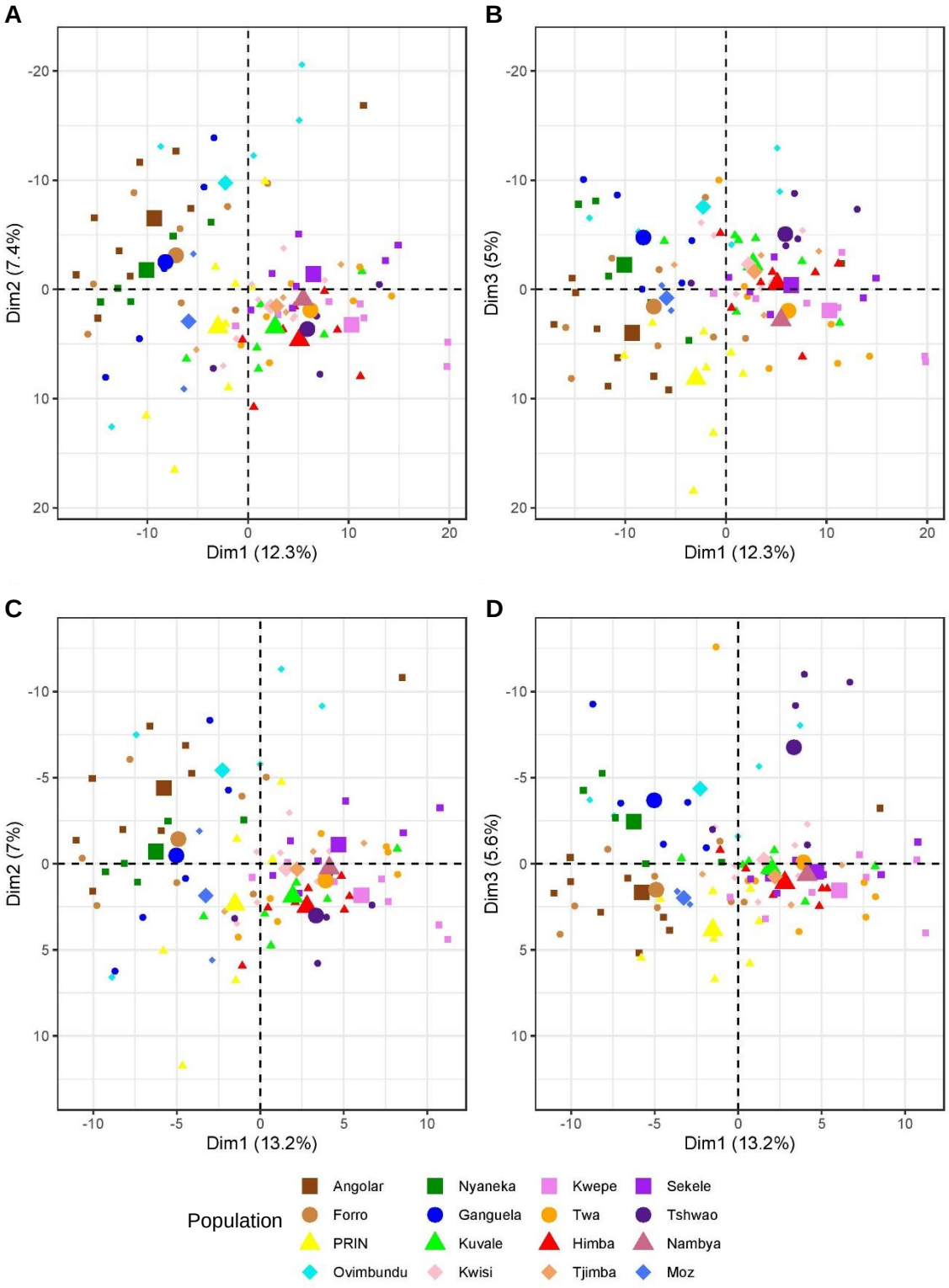


Figure 10 | Principal component analysis based on microbial species (A and B) and genera (C and D) data from 95 African individuals. A and C represent PC1 and PC2 whereas B and D represent PC1 and PC3. Individuals are represented by a specific combination of symbol and colour representative of the population. Population midpoints are indicated with larger symbols.

Besides the positioning of the individuals, PCA also allows us to see the positioning of the variables (here species and genera) by means of a loading plot of the PCA. Variables are plotted in accordance to their weight in each component, and those plotted in the same direction are positively correlated whereas variables plotted in opposite direction are negatively correlated. The position of individuals in the PCA is also correlated with the direction of the variables, therefore, individuals plotted in the same direction of a determined variable, are enriched in that variable (species or genera).

Figure 11A shows the 20 species with the greatest contribution to PC1-PC2, most of the species were positioned concordant with the cluster of non-agropastoral populations (i.e. they are in the bottom right side of the graph). This pattern indicates that most of the differences between populations are driven by species enriched in non-agropastoralists and with low frequency in agropastoralists. Overall the species with more contribution to PC1-PC2 are *Streptococcus parasanguinis*, *Streptococcus infantis*, *Atopobium sp.*, *Streptococcus peroris and Veillonella dispar*.

Figure 11B shows the 20 genera with the greatest contribution to PC1-PC2, the positioning of genera in relation to that of the species was more structured. Most of the genera (16) are on the most extreme positive side of PC1, thus, are enriched in the non-agropastoralists cluster. In contrast, the genus *Eikenella* is enriched in the agropastoralists cluster. The genera *Gemella*, *Granulicatella* and *Streptococcus* are positively correlated with PC2, with the highest abundance present in Príncipe, and the lower in Ovimbundu and Angolar. Overall, the genera with more contribution to PC1-PC2 are *Atopobium*, *Eubacterium*, *Campylobacter*, *Dialister* and *Lancefieldella*.

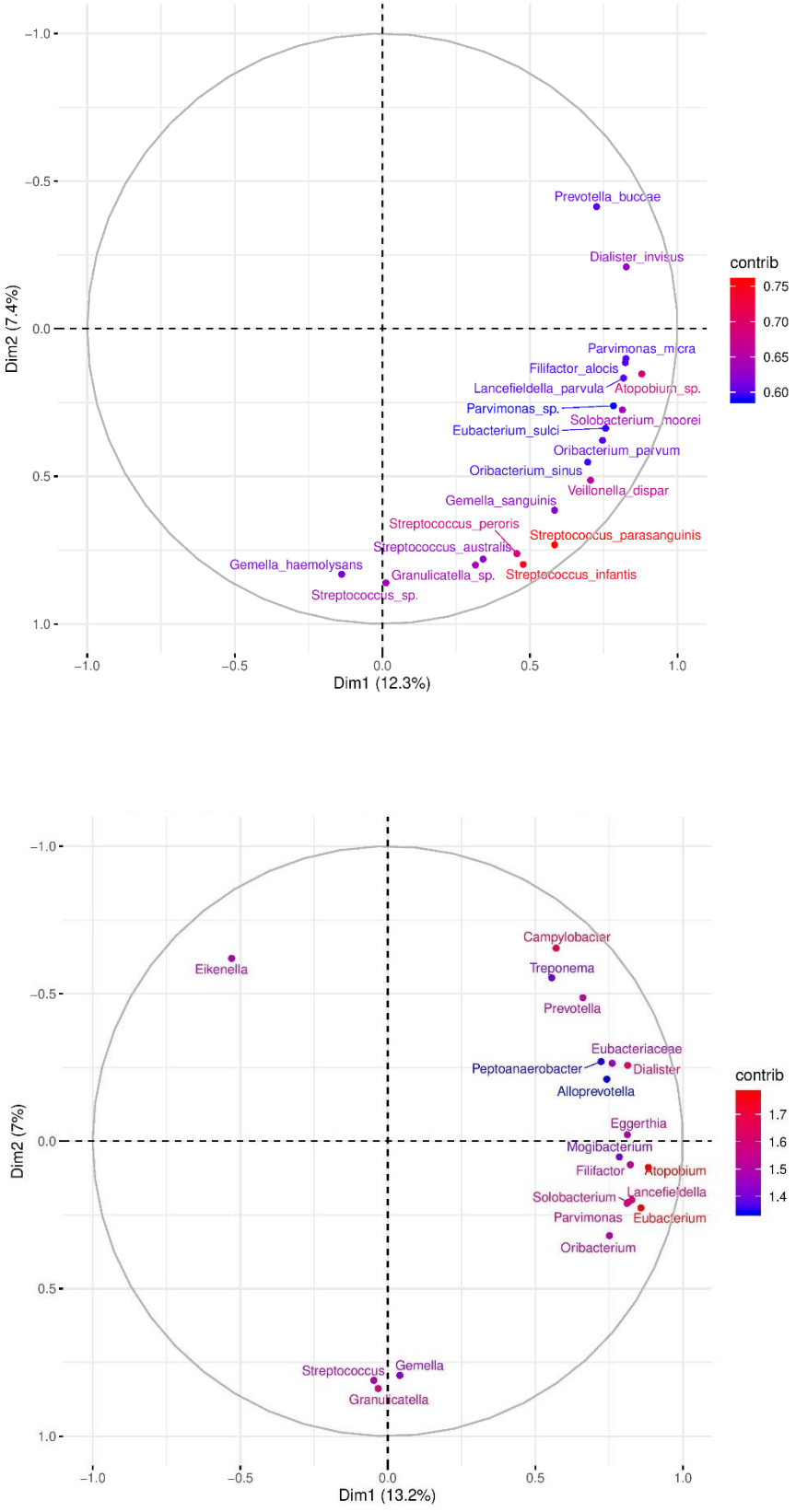


Figure 11 | Loading plots (positioning of the variables species and genera) of the principal component analysis based on microbial species (A) and genera (B) data from 95 African individuals. For both PCA the 20 species/genera with the greatest contribution to PC1-PC2 are represented. Species/Genera are coloured by contribution as indicated by the legend.

We further explored principal component analysis (both with species and genera data) with the southern Africa dataset (Angola, Zimbabwe and Mozambique) without considering those from São Tomé and Príncipe (Figure S6 and S7, Appendix), and the Angola dataset (Figure S8 and S9, Appendix).

The relative position of the populations in these PCAs was generally maintained. The main difference was observed on the PCA based on species frequencies of southern Africa individuals, in which PC3 separates the Tshwao from the other groups. This separation was already seen in the PCA with all individuals, but only when accounting for genera frequencies.

For the PCA considering only southern Africa individuals, the taxa with more contribution to the first two components were, in the case of the species, *Streptococcus parasanguinis*, *Granulicatella sp.*, *Granulicatella adiacens*, *Rothia sp. and Veillonella dispar*, and in the case of the genera, *Atopobium*, *Lancefieldella*, *Solobacterium*, *Eubacterium* and *Oribacterium* (Figure S7 A and B, Appendix). For the PCA considering only Angolan individuals, the taxa with more contribution to the first two components were, in the case of the species, *Streptococcus parasanguinis*, *Rothia sp.*, *Treponema socranskii*, *Veillonella dispar* and *Atopobium sp.*, and in the case of the genera, *Lancefieldella*, *Atopobium*, *Solobacterium*, *Oribacterium* and *Rothia* (Figure S9, A and B, Appendix).

In order to further explore the differences among individuals we performed another ordination method, a non-metric multidimensional scaling (NMDS). Both methods aim to minimize the dimensions of the data, but PCA preserves the covariance and MDS preserves the distance between individuals. NMDS was performed from the Bray–Curtis dissimilarity distance matrix for both species and genera data with all the individuals, with southern African individuals and with Angolan individuals. We used the option of displaying the population's polygons, created by connecting individuals of the same population. In NMDS1 and NMDS2 we observe two clusters, one with low dispersion of individuals, which comprehend mainly the non-agropastoral groups and the other with high dispersion of individuals, which comprehend the São Tomé and Príncipe and Angolan agropastoralists. The two clusters show little overlap and are better discriminated while based on species frequencies data. The two NMDS accounting only for southern Africa individuals and only for Angolan individuals display a more striking separation of the two clusters. The NMDS results are fairly consistent with those of PCA (Figure 12; Figure S10 and S11, Appendix).

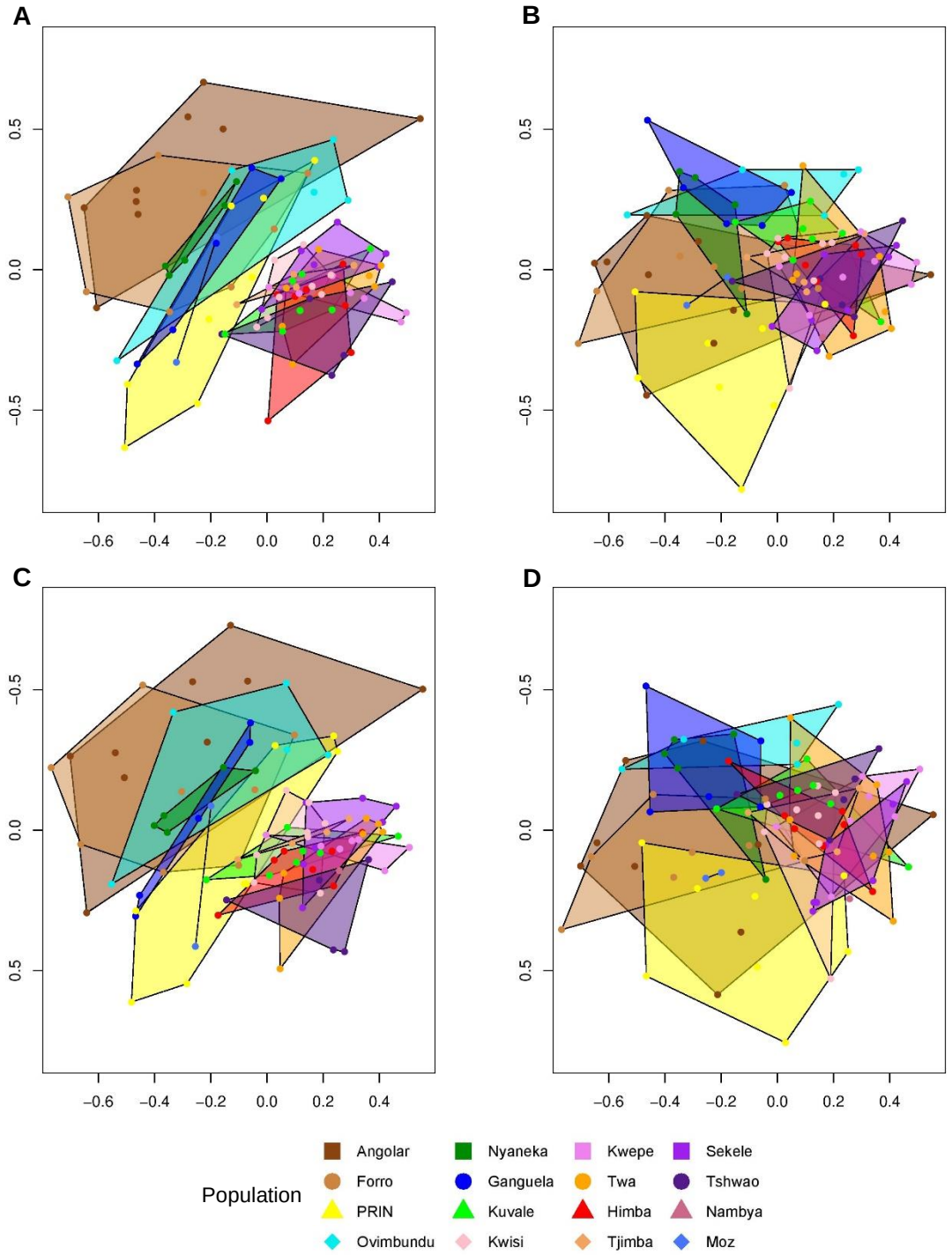


Figure 12 | Non-metric multidimensional scaling based on microbial species (A and B) and genera (C and D) data from 95 African individuals (in both cases a three-dimensional NMDS was built). A and C represent dimensions 1 and 2 of the NMDS, while B and D represent dimensions 1 and 3. On the species based NMDS (A and B) the stress level was 0.124 whereas on the genera based NMDS (C and D) the stress level was 0.144. In order to better understand the positioning of the populations in the low-dimension space, the individuals with more extreme positions in each population where connected by lines forming population polygons represented by a specific colour as indicated by the legend.

To visualize how the oral microorganisms are distributed across individuals we built heatmaps of species and genera abundance, which are shown in Figure S12 and S13.

We wanted to focus on the species and genera with more contribution for the PCA, so we built heatmaps considering the 20 species/genera with more weight for the three datasets: All individuals, Southern African individuals and Angolan individuals. We have also performed a hierarchical clustering of the individuals in order to understand which ones present the most similar oral microbiomes. The hierarchical cluster analysis supports the separation of individuals in two clusters that largely correspond to the agropastoralists and the non-agropastoralists as already observed in the PCA and NMDS results. It is important to mention that the Nambya individual fall inside the non-agropastoralists cluster. In the hierarchical cluster analysis while accounting for all the individuals and using species data, one individual from Mozambique and 3 individuals from Príncipe also fall within the non-agropastoralists (Figure 13). When we consider genera data, all individuals from Mozambique and Príncipe, plus two individuals from Angolar and three Forro fall within the cluster of non-agropastoralists (Figure 14). In the heatmaps based on the Southern Africa and Angola datasets, the hierarchical clustering of individuals separates more clearly agropastoral from non-agropastoral groups, except in the Southern Africa case, where, besides Nambya, one individual from Mozambique also falls within the non-agropastoral group (Figure S14 – S17, Appendix).

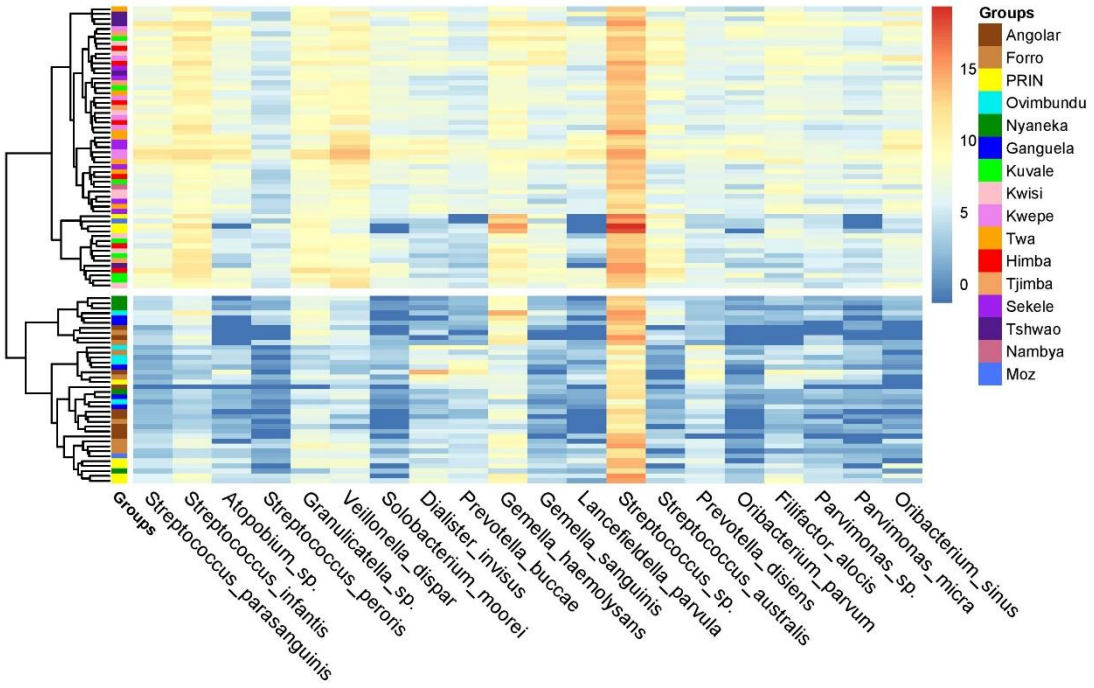


Figure 13 | Heatmap of species abundance in 95 African individuals. Only the 20 species with the greatest contribution to principal component 1 and 2 of the respective PCA are shown. Individuals (rows) are coloured in relation to their population.

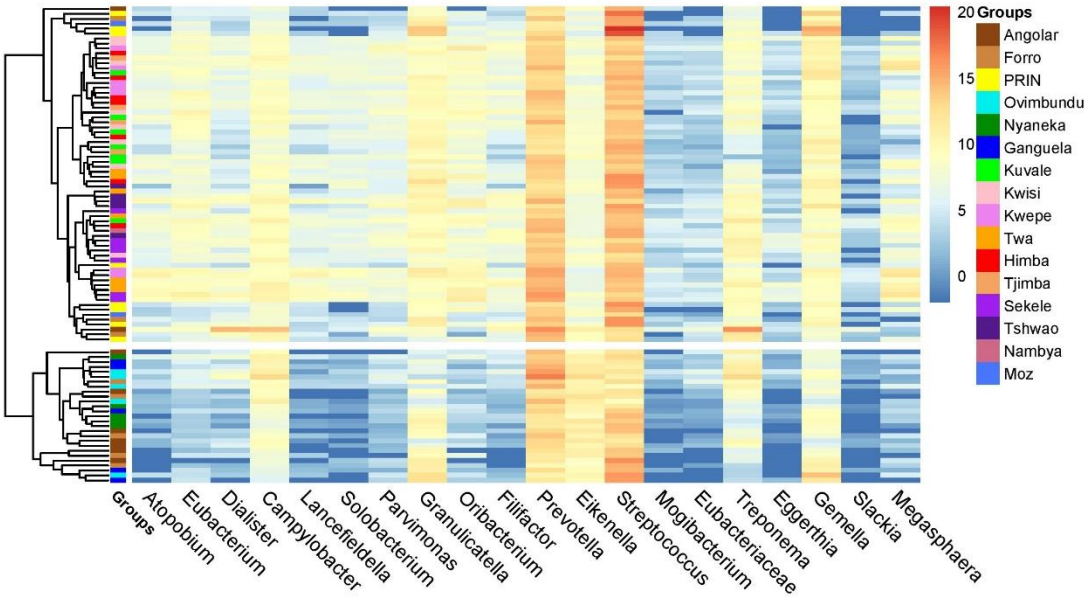


Figure 14 | Heatmap of genera abundance in 95 African individuals. Only the 20 genera with the greatest contribution to principal component 1 and 2 of the respective PCA are shown. Individuals (rows) are coloured in relation to their population.

3.4 Factors shaping microbial profiles in African populations

In order to test if the oral microbiome composition varies among groups defined upon populations characteristics (e.g. subsistence patterns), we performed a PERMANOVA analysis (Table 2). In cases were the PERMANOVA results were significant, we also evaluated the dispersion of the groups with an Anova analysis in order to see if the differences solely resulted from the groups' centroids. We compared groups of populations based on type of collected sample, languages, geography (countries) and subsistence patterns, as described in Table 1. Taking into account the results obtained in the other analysis, namely PCA, NMDS and hierarchical clustering, we also performed the analysis referent to subsistence pattern by comparing agropastoralists with non-agropastoralists, with the original classifications of each population and by including Nambya in the non-agropastoral group. All the PERMANOVA and dispersion comparisons were significant. However, the comparison in which PERMANOVA returned the higher pseudo-F value, a statistic referent to the strength of the test, was Agropastoralists vs Non-Agropastoralists (including Nambya) (Pseudo-F = 21.888), in which 19% of the total variance could be explained by differences between these two categories.

Table 2 | Results from PERMANOVA analysis. Grouping criteria defined as in table 1.

	Anova results from clusters dispersion		PERMANOVA results		
Grouping criteria	p-values	F-value	p-value	R2	Pseudo-F value
Sample type	4.487e-06	4.5084	0,0001	0.07715	7.7752
Language	0.000743	6.1518	0,0001	0.15019	5.3611
Country	0.01937	3.4693	0,0001	0.1362	4.7828
Subsistence	2.412e-10	20.787	0,0001	0.20267	7.7104
Subsistence op. 2	1.809e-10	21.117	0,0001	0.20541	7.8416
Agro vs Non-Agropastoralists	1.421e-10	52.067	0,0001	0.18235	20.741
Agro vs Non-Agropastoralists op. 2	2.363e-10	50.518	0,0001	0.19051	21.888

Subsistence op. 2 - Equal to “Subsistence” but considering Tshwao as foragers.

Agro vs Non-Agropastoralists op. 2 - Equal to “Agro vs Non-Agropastoralists” but considering Nambya as Non-Agropastoralists.

4. Discussion

In this work we characterized the oral microbiome of 95 individuals from 16 human populations from Africa by using the portion of non-human reads obtained from an Exome Capture Sequencing of DNA from saliva samples and cheek scraps.

We observed slightly lower percentage of non-human reads in the cheek scraps (1.64%) than in saliva samples (2.43%). Kidd et al., (2014) used this approach on saliva samples and obtained a slightly higher percentage than us, 5.2%. This difference could be due to the use of a different enrichment capture kit. Nevertheless, we were able to characterize the oral microbiome and record patterns of species and genera distribution consistent with the ones obtained from other approaches. In that sense, we found that a small fraction of microorganisms recruited most of the reads. Namely, the 20 most abundant species recruited 68.7% of all reads while the 100 most abundant species recruited 95%. We identified the most abundant genera across all studied populations and these comprised *Neisseria*, with 21% of the reads, *Streptococcus* (17.3%), *Prevotella* (12%), *Rothia* (7.5%) and *Porphyromonas* (6.3%), which is consistent with other studies (Nasidze et al., 2009; Li et al., 2014; Kidd et al., 2014).

With respect to the number of species and genera detected in the different populations studied we found a considerably lower number of taxa in the populations from

São Tomé and Príncipe and Mozambique. Two factors could explain these differences; first, 24 out of the 25 individuals from São Tomé and Príncipe were sequenced in a different run for which a lower number of reads was obtained. However, taking into account that one individual from São Tomé, and the two from Mozambique were sequenced jointly with all remaining samples, but they also present a considerably lower number of taxa, we speculate that the observed differences are mainly caused by the type of collected sample rather than a batch effect of the sequencing runs. Secondly, these populations are the only ones whose DNA was extracted from cheek scraps, and therefore, it may reflect that the microbial community present in the cheeks is less diverse and less representative of the other oral sites than the microbial community of the saliva, as in fact was already reported (Zaura et al., 2009).

In any case, to account for the effects of the unequal number of sequence reads obtained per individual which was correlated with the number of taxa identified, these analyses should be repeated using a rarefaction of the reads i.e. for each individual we will randomly subsample a number of reads equal to the individual with fewer reads.

In comparison with previous studies that used the amplicon sequencing of 16S rDNA gene (Nasidze et al., 2009; Nasidze et al., 2011; Li et al., 2014), the number of taxa we detected was considerably higher, which might reveal to be an advantage of the Exome capture sequencing approach. However, we should not exclude the possibility that the actual number of microorganisms present in the samples is lower than what we estimated, since as Kidd et al., (2014) mentioned, some microorganisms present genomic regions of high identity with the human genome and could be falsely detected by low-quality human reads.

We measured the diversity of the oral microbiome on the different populations and generally found lower levels of intra-individual diversity (alpha diversity) and higher levels of inter-individual diversity (beta diversity) in most agropastoral groups in relation to groups having other subsistence modes. Interestingly, the only hunter-gatherers of our study, the Sekele, presented the highest values of alpha diversity and one of the lowest values of beta diversity. These results are consistent with what was found in other studies, such as, Nasidze et al., (2011), which compared the former hunter-gatherers Batwa Pygmies from Uganda with agricultural groups from Sierra Leone and the Democratic Republic of Congo, and Lassalle et al., (2018), which compared hunter-gatherer and traditional farmer populations from the Philippines.

One of the most striking results was the separation between agropastoral and non-agropastoral groups observed in the ordination and cluster analysis. This separation is

more evident at the species level than at the genus level, factor that enhances the importance of applying shotgun sequencing over amplicon sequencing methods on the study of the oral microbiome, to allow microbial identification at the species level.

In both PCA and NMDS results, the agropastoral individuals are highly dispersed while the non-agropastoral ones are more clustered. These dispersion patterns are concordant with the higher inter-individual diversity values in agropastoralists than in non-agropastoralists. The fact that the cluster of agropastoralists also includes the three agropastoral populations from Angola, which samples were obtained through saliva, suggests that the differences between these two clusters cannot only be due to differences in the microbiome of cheek scraps and saliva. It is worth mentioning that the Nambya from Zimbabwe, which were classified as agropastoralists fall inside the non-agropastoral cluster.

When focusing on Southwestern Angola, it is striking that no differences in the oral microbiome profile were observed between the forager Khoisan-speaking Sekele, and the Bantu-speaking pastoral and peripatetic populations, and that the larger difference was between all these populations versus the agropastoral ones.

We tested whether there were statistically significant differences between the microbial profiles of the individuals according to diverse characteristics of the populations studied with PERMANOVA analysis. Although we found significant differences according to all comparisons performed, the individuals were better separated in agropastoralists versus non-agropastoralists. It is important to highlight that the groups compared present also different dispersions, which can result in significant values for PERMANOVA. Whereby we cannot conclude that the groups have in fact different compositions. Nevertheless, taking into account the results from the ordination and cluster analysis, it is reasonable to state that the agropastoral and the non-agropastoral groups present different compositions of the oral microbiomes, and that the significant value obtained with the PERMANOVA test would not only result from the groups different dispersions.

One of the reasons that could be responsible for the differences between these two groups is diet as has been documented (Nasidze et al., 2011; Nam et al., 2011; Schnorr et al., 2014). Agropastoralists probably present a higher carbohydrate intake coming from the crops they grow compared to hunter-gatherers, which likely present a higher animal protein intake derived from game food. However, since we lack reasonable information about differences in nutrient intake between these populations, especially referent to the Angolan peripatetics (Kwisi, Twa and Tjimba) is not possible to conclude whether the observed differences are reflective of their different diets or not.

Another plausible reason for the observed differences in the oral microbiome of agropastoral and non-agropastoral groups could be related with a socio-economic status. Agropastoralist's primary economic activity are growing crops and raising livestock, which represent a higher profit activity than foraging or being a pastoralist. Thus, when focusing on Angola, the agropastoralists Nyaneka, Ganguela and Ovimbundu are wealthier and have a higher status than individuals with other lifestyles. In addition, the economic status could be related with the level of hygiene, and so, higher in agropastoralists. Kidd et al., (2014) found that, the oral microbiome of KhoeSan presented several known pathogens among the most abundant taxa, what was not true among healthy Americans. The authors suggested that this could result from a "limited access to dental care, antibiotics and/or absence of water fluoridation among the KhoeSan". This hypothesis could explain the differences observed in our populations, but should be studied in more detail. Other studies have speculated about the influence of the oral hygiene in the oral microbiome. Clement et al., (2015) proposed that the similar levels of diversity observed between the Yanomami and the U.S. individuals could result from the level of oral hygiene and from the habit of chewing of tobacco, however, in terms of composition, the oral microbiomes from these two populations was different.

In order to obtain more accurate conclusions about the factors responsible for the differentiation of the oral microbiomes in agropastoral and non-agropastoral groups, it could be useful to investigate in which biological processes the species or genera with different abundances between agropastoral and non-agropastoral groups are involved. Likewise, the pathogen load of those populations might give clues about their hygienic levels. Finally, further analysis comparing the microbiome profiles of the populations here studied with those from other hunter-gatherer and agropastoral groups (Nasidze et al., 2011), as well as from populations having a western lifestyle (Lassalle et al., 2018) would be useful to understand the pattern here observed.

In conclusion, we corroborate that the Human Exome Capture Sequencing approach applied to oral samples allows a faithful characterization of the human oral microbiome. Through this approach we conclude that several African populations present distinct oral microbiome profiles, which mainly discriminate agropastoral from non-agropastoral groups, beyond this, as reported in previous studies, the agropastoral individuals present the less diverse oral microbiomes. These differences could result from the diet and other conditions, which are likely related to differences in the economic-status, like the levels of hygiene, although more studies are needed. Taking into account that the diverse populations analysed are from an understudied region, our study provides valuable

information to a more complete comprehension of the global human oral microbiome composition.

5. References

Adler, C. J., Dobney, K., Weyrich, L. S., Kaidonis, J., Walker, A. W., Haak, W., Bradshaw, C. J. A., Townsend, G., Sołtysiak, A., Alt, K. W., Parkhill, J., & Cooper, A. (2013). Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature genetics*, *45*(4), 450.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403-410.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral ecology*, *26*(1), 32-46.

Blekhman, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., Spector, T. D., Keinan, A., Ley, R. E., Gevers, D., & Clark, A. G. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome biology*, *16*(1), 191.

Bray, J. R., & Curtis, J. T. (1957). An ordination of the upland forest communities of south ern Wisconsin. *Ecological monographs*, *27*(4), 325-349.

Chen, H., & Jiang, W. (2014). Application of high-throughput sequencing in understanding human oral microbiome related with health and disease. *Frontiers in microbiology*, *5*, 508.

Clemente, J. C., Pehrsson, E. C., Blaser, M. J., Sandhu, K., Gao, Z., Wang, B., Magris, M., Hidalgo, G., Contreras, M., Noya-Alarcón, Ó., Lander, O., McDonald, J., Cox, M., Walter, J., Oh, P. L., Ruiz, J. F., Rodriguez, S., Shen, N., Song, S, J., Metcalf, J., Knight, R., Dantas, G., & Dominguez-Bello, M. G. (2015). The microbiome of uncontacted Amerindians. *Science* advances, *1*(3), e1500183.

Demmitt, B. A., Corley, R. P., Huibregtse, B. M., Keller, M. C., Hewitt, J. K., McQueen, M. B., Knight, R., McDermott, I., & Krauter, K. S. (2017). Genetic influences on the human oral microbiome. *BMC genomics*, *18*(1), 659.

Dray, S., & Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, *22*(4), 1-20.

Eren, A. M., Borisy, G. G., Huse, S. M., & Welch, J. L. M. (2014). Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Sciences*, *111*(28), E2875-E2884.

Fox, J., & Weisberg, S. (2019). An {R} Companion to Applied Regression, Third Edition. T housand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Com panion/

Grassl, N., Kulak, N. A., Pichler, G., Geyer, P. E., Jung, J., Schubert, S., Sinitcyn, P., Cox, J., & Mann, M. (2016). Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. *Genome medicine*, *8*(1), 44.

Gupta, V. K., Paul, S., & Dutta, C. (2017). Geography, ethnicity or subsistence-specific variations in human microbiome composition and diversity. *Frontiers in microbiology*, *8*, 1162.

Joshi, N. A., & Fass, J. N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at: https://github.com/najoshi/sickle.

Kassambara, A., & Mundt, F. (2017). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. https://CRAN.R-project.org/package=factoextra

Kidd, J. M., Sharpton, T. J., Bobo, D., Norman, P. J., Martin, A. R., Carpenter, M. L., Sikora, M., Gignoux, C. R., Nemat-Gorgani, N., Adams, A., & Guadalupe, M. (2014). Exome capture from saliva produces high quality genomic and metagenomic data. *BMC genomics*, *15*(1), 262.

Kolde, R. (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap

Lassalle, F., Spagnoletti, M., Fumagalli, M., Shaw, L., Dyble, M., Walker, C., Thomas, M. G., Migliano, A. B., & Balloux, F. (2018). Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Molecular ecology*, *27*(1), 182-195.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv preprint arXiv:1303.3997*. Available at: http://arxiv.org/abs/1303.3997

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079.

Li, J., Quinque, D., Horz, H.-P., Li, M., Rzhetskaya, M., Raff, J. A., Hayes, M. G., & Stoneking, M. (2014). Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. *BMC microbiology*, *14*(1), 316.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550.

Lu, M., Xuan, S., & Wang, Z. (2019). Oral microbiota: A new view of body health. *Food Science and Human Wellness*, *8*(1), 8-15.

Mager, D. L., Ximenez-Fyvie, L. A., Haffajee, A. D., & Socransky, S. S. (2003). Distribution of selected bacterial species on intraoral surfaces. *Journal of clinical periodontology*, *30*(7), 644-654.

Mason, M. R., Nagaraja, H. N., Camerlengo, T., Joshi, V., & Kumar, P. S. (2013). Deep sequencing identifies ethnicity-specific bacterial signatures in the oral microbiome. *PloS one*, *8*(10), e77287.

McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, *10*(4), e1003531.

Methé, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., ... & Chinwalla, A. T. (2012). A framework for human microbiome research. *nature*, *486*(7402), 215.

Murtagh, F., & Legendre, P. (2014). Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification*, *31*(3), 274-295.

Nakano, Y., Suzuki, N., & Kuwata, F. (2018). Predicting oral malodour based on the microbiota in saliva samples using a deep learning approach. *BMC oral health*, *18*(1), 128.

Nam, Y. D., Jung, M. J., Roh, S. W., Kim, M. S., & Bae, J. W. (2011). Comparative analys is of Korean human gut microbiota by barcoded pyrosequencing. *PloS one*, *6*(7), e2 2109.

Nasidze, I., Li, J., Quinque, D., Tang, K., & Stoneking, M. (2009). Global diversity in the human salivary microbiome. *Genome research*, *19*(4), 636-643.

Nasidze, I., Li, J., Schroeder, R., Creasey, J. L., Li, M., & Stoneking, M. (2011). High diversity of the saliva microbiome in Batwa Pygmies. *PloS one*, *6*(8), e23352.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., Szoecs, E., & Wagner, H. (2019). vegan: Community Ecology Package. R pa ckage version 2.5-6. https://CRAN.R-project.org/package=vegan

Osborne, J. (2002). Notes on the use of data transformations. *Practical assessment, research and evaluation*, *9*(1), 42-50.

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., & Konstantinidis, K. T. (2014). Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PloS one*, *9*(4), e93827.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841-842.

Quinque, D., Kittler, R., Kayser, M., Stoneking, M., & Nasidze, I. (2006). Evaluation of saliva as a source of human DNA for population and association studies. *Analytical biochemistry*, *353*(2), 272-277.

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and biophysical research communications*, *469*(4), 967-977.

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y.-H., Falcón, L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R., Strausberg, R. L., Nealson, K., Friedman, R., Frazier, M., & Venter, J. C. (2007). The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology*, *5*(3), e77.

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, *27*(6), 863-864.

Schnorr, S. L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., Fiori, J., Gotti, R., Bellis, G. D., Luiselli, D., Brigidi, P., Mabulla, A., Marlowe, F., Henry A. G., & Crittenden A. N. (2014). Gut microbiome of the Hadza hunter-gatherers. *Nature communications*, *5*, 3654.

Schnorr, S. L., Sankaranarayanan, K., Lewis Jr, C. M., & Warinner, C. (2016). Insights into human evolution from ancient and contemporary microbiome studies. *Current opinion in genetics & development*, *41*, 14-26.

Sender, R., Fuchs, S., & Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS biology*, *14*(8), e1002533.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical jo urnal*, *27*(3), 379-423.

Takeshita, T., Matsuo, K., Furuta, M., Shibata, Y., Fukami, K., Shimazaki, Y., Akifusa, S., Han, D.-H., Kim, H.-D., Yokoyama, T., Ninomiya, T., Kiyohara, Y., & Yoshihisa Yamashita, Y. (2014). Distinct composition of the oral indigenous microbiota in South Korean and Japanese adults. *Scientific reports*, *4*, 6990.

Team, R. (2016). RStudio: Integrated development for R [Computer software]. *URL http://www.rstudio.com/.Boston, MA*: RStudio, Inc.

Team, R. C. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012. *URL http://www.R-project. org*.

Wade, W. G. (2013). The oral microbiome in health and disease. *Pharmacological research*, *69*(1), 137-143.

Zaura, E., Keijser, B. J., Huse, S. M., & Crielaard, W. (2009). Defining the healthy "core microbiome" of oral microbial communities. *BMC microbiology*, *9*(1), 259.

6. Appendix

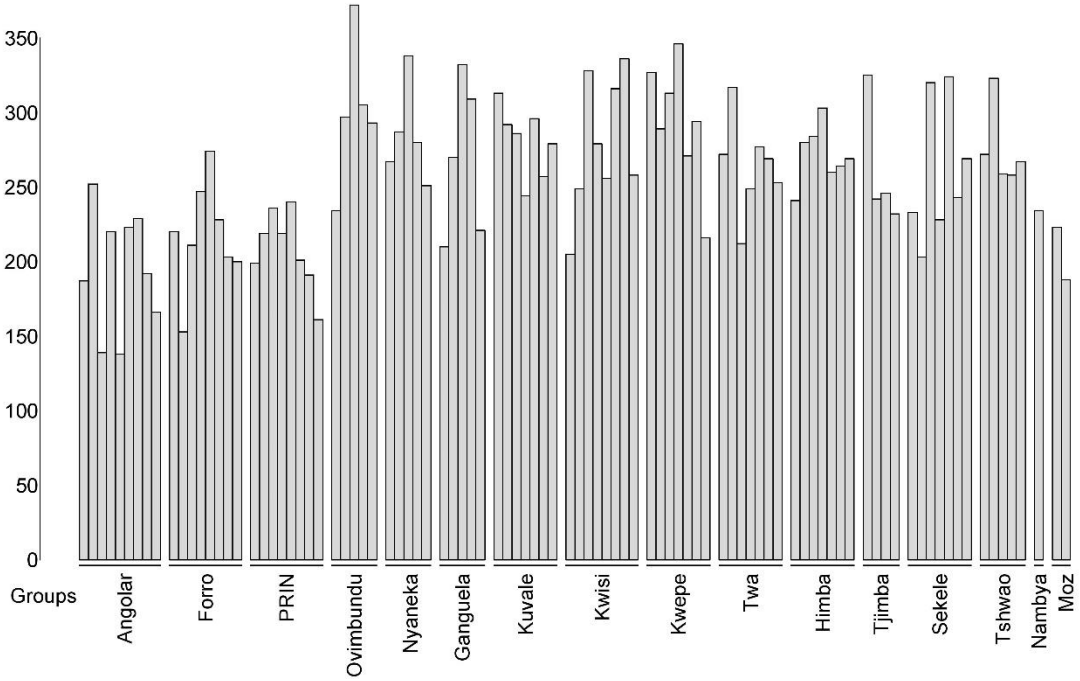


Figure S1 | Bar plots showing the number of microbial species identified for the individuals of each population.

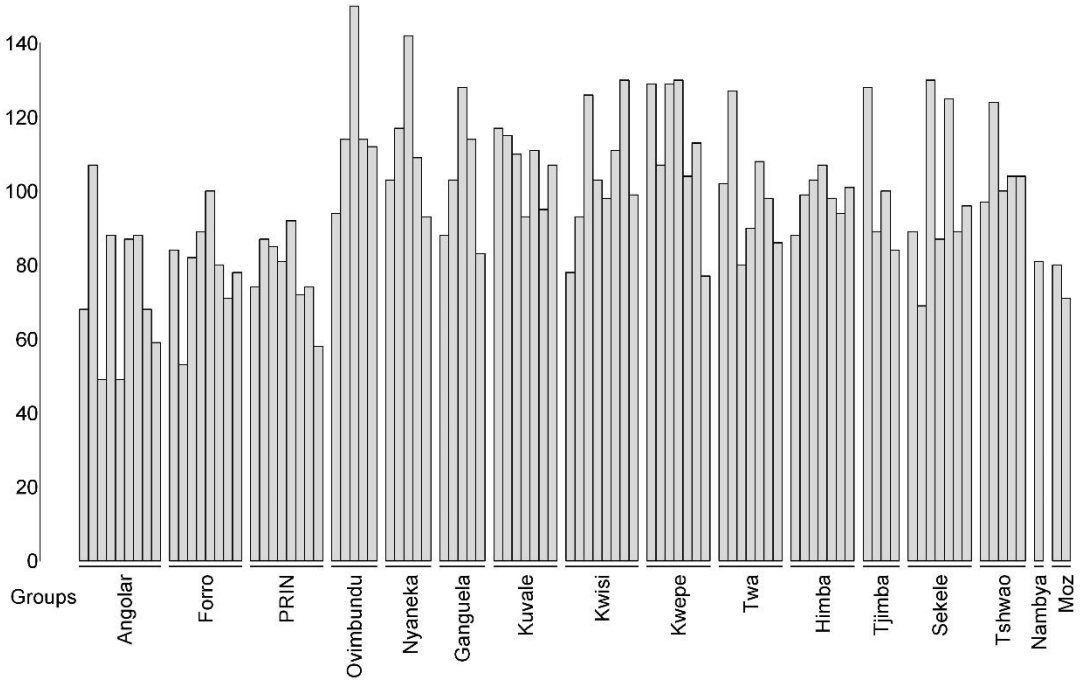


Figure S2 | Bar plots showing the number of microbial genera identified for the individuals of each population.

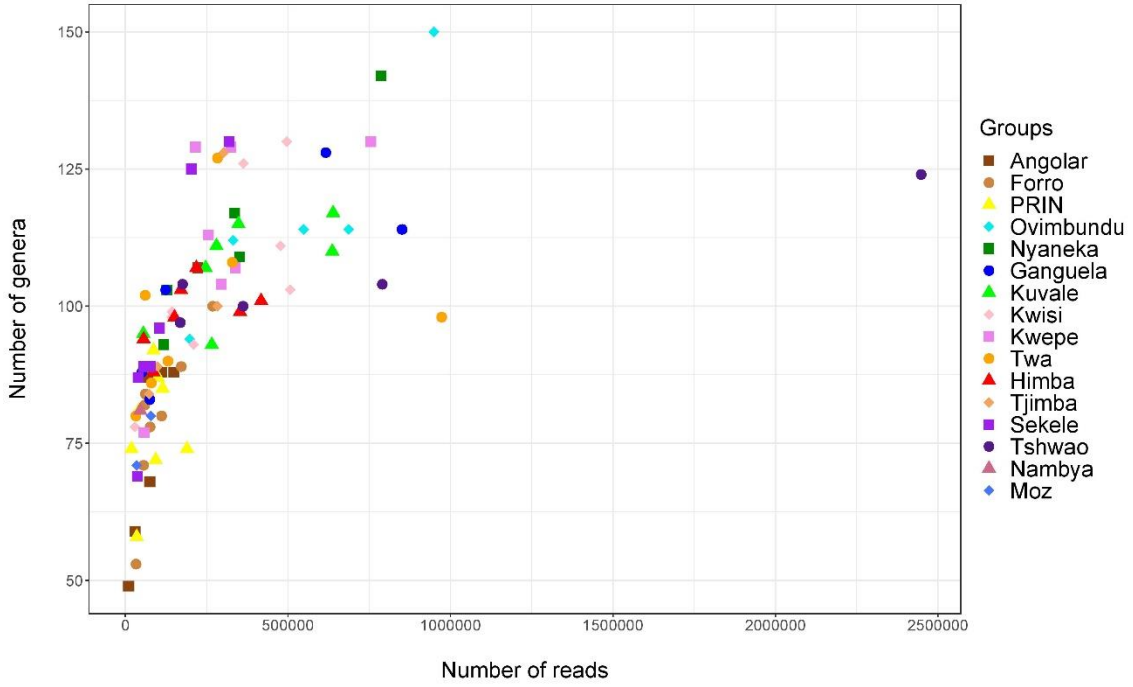


Figure S3 | Relation between the number of reads obtained for each individual and the correspondent number of microbial genera identified. Spearman correlation between these two variables: r_s = 0.851 p-values < 2.2e-16.

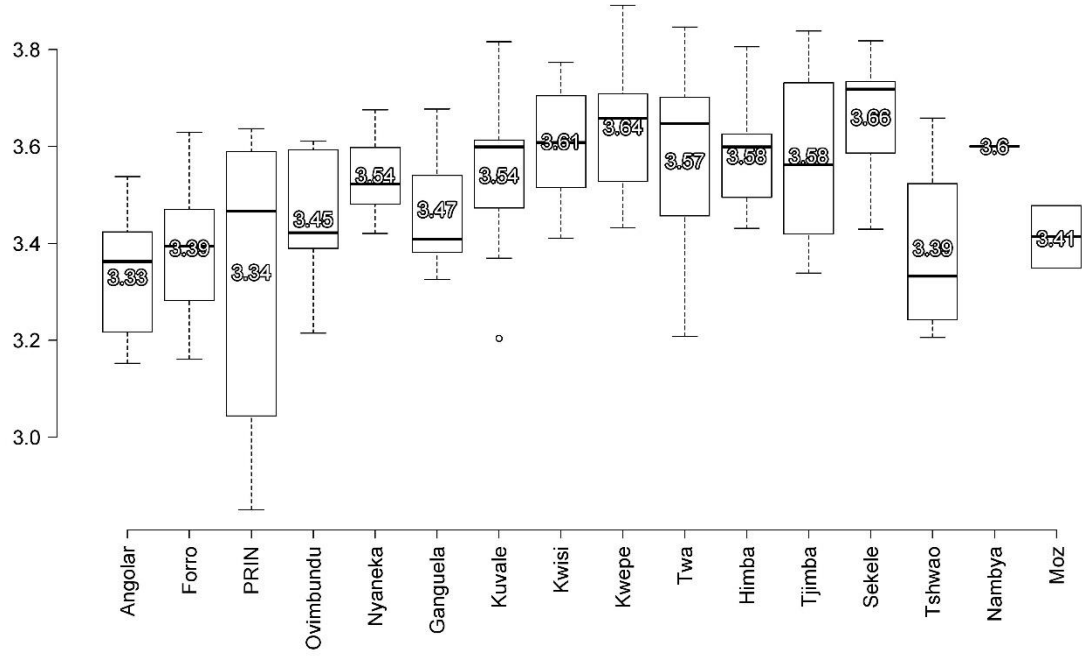


Figure S4 | Box plots showing the distribution of alpha diversity values (Shannon index) calculated for the individuals of each population at the genus level. Numbers inside box plots correspond to populations mean values.

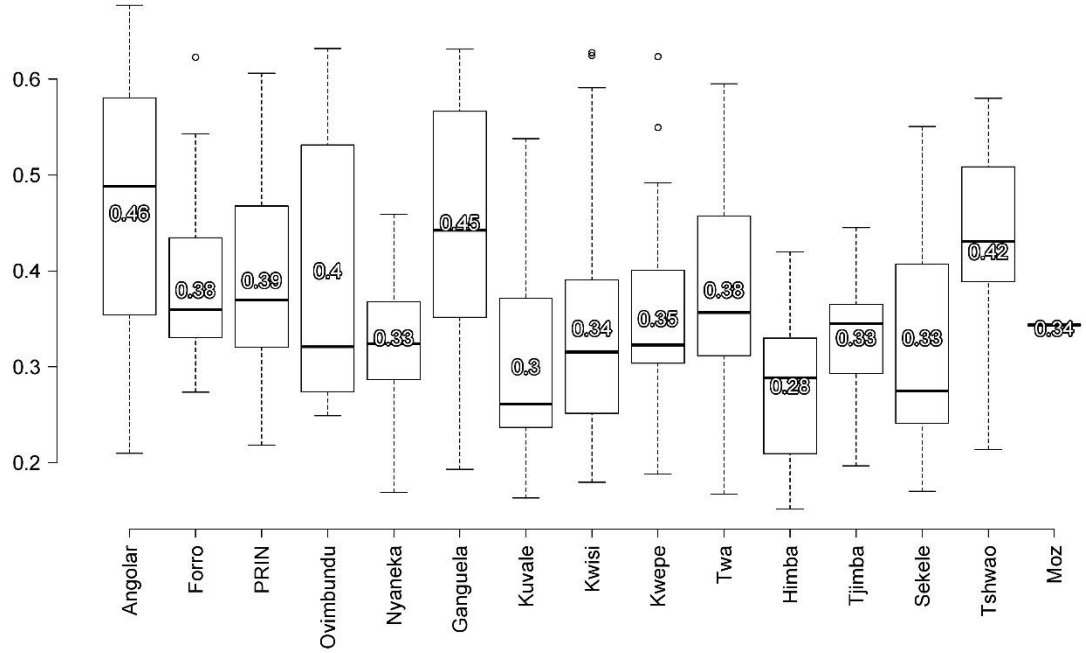


Figure S5 | Box plots showing the distribution of beta diversity values (Bray–Curtis dissimilarity) calculated between individuals of each population at the genus level. Numbers inside box plots correspond to populations mean values.

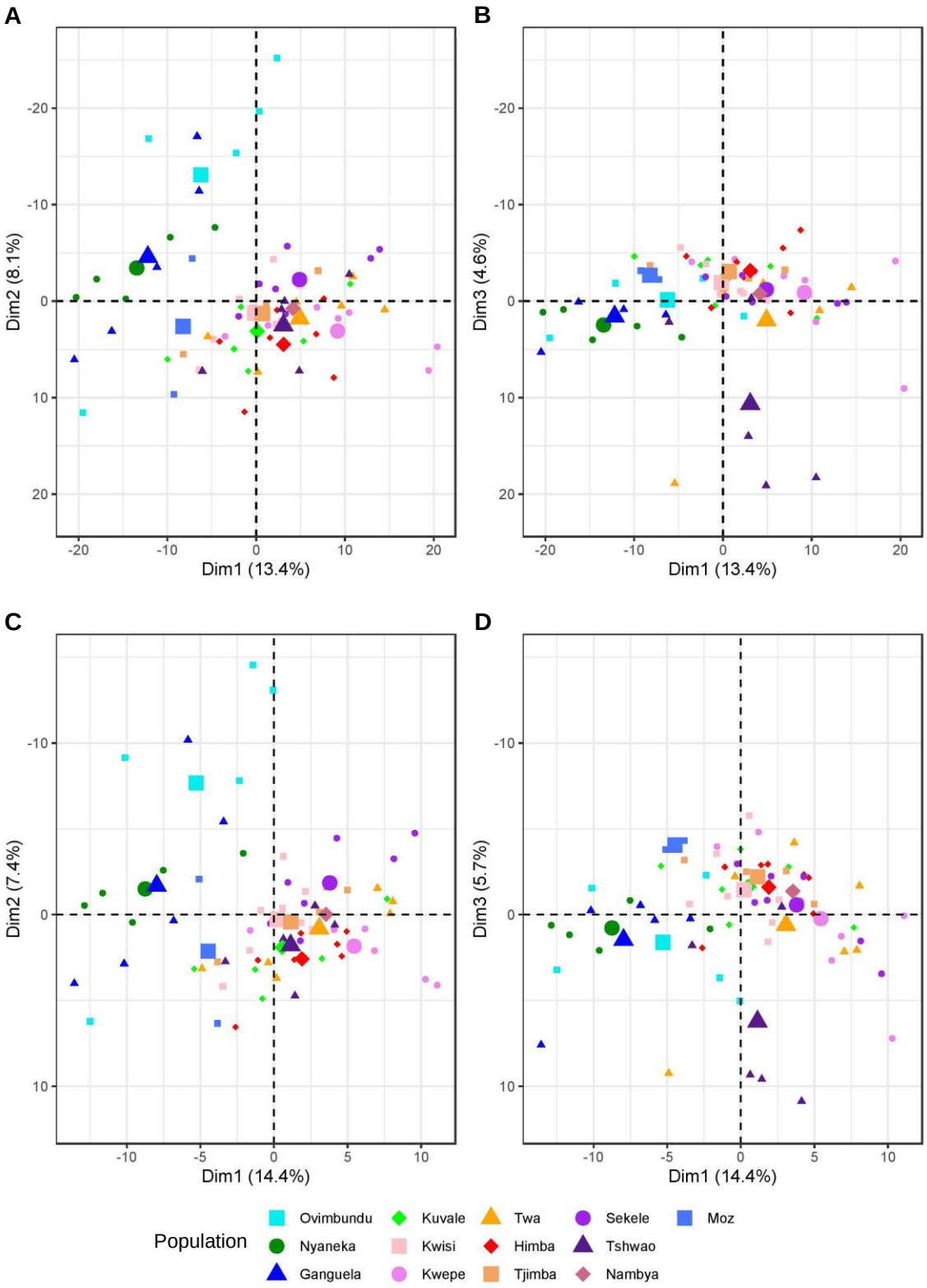


Figure S6 | Principal component analysis based on microbial species (A and B) and genera (C and D) data from 70 Southern African individuals. A and C represent PC1 and PC2 whereas B and D represent PC1 and PC3. Individuals are represented by a specific combination of symbol and colour representative of the population. Population midpoints are indicated with larger symbols.

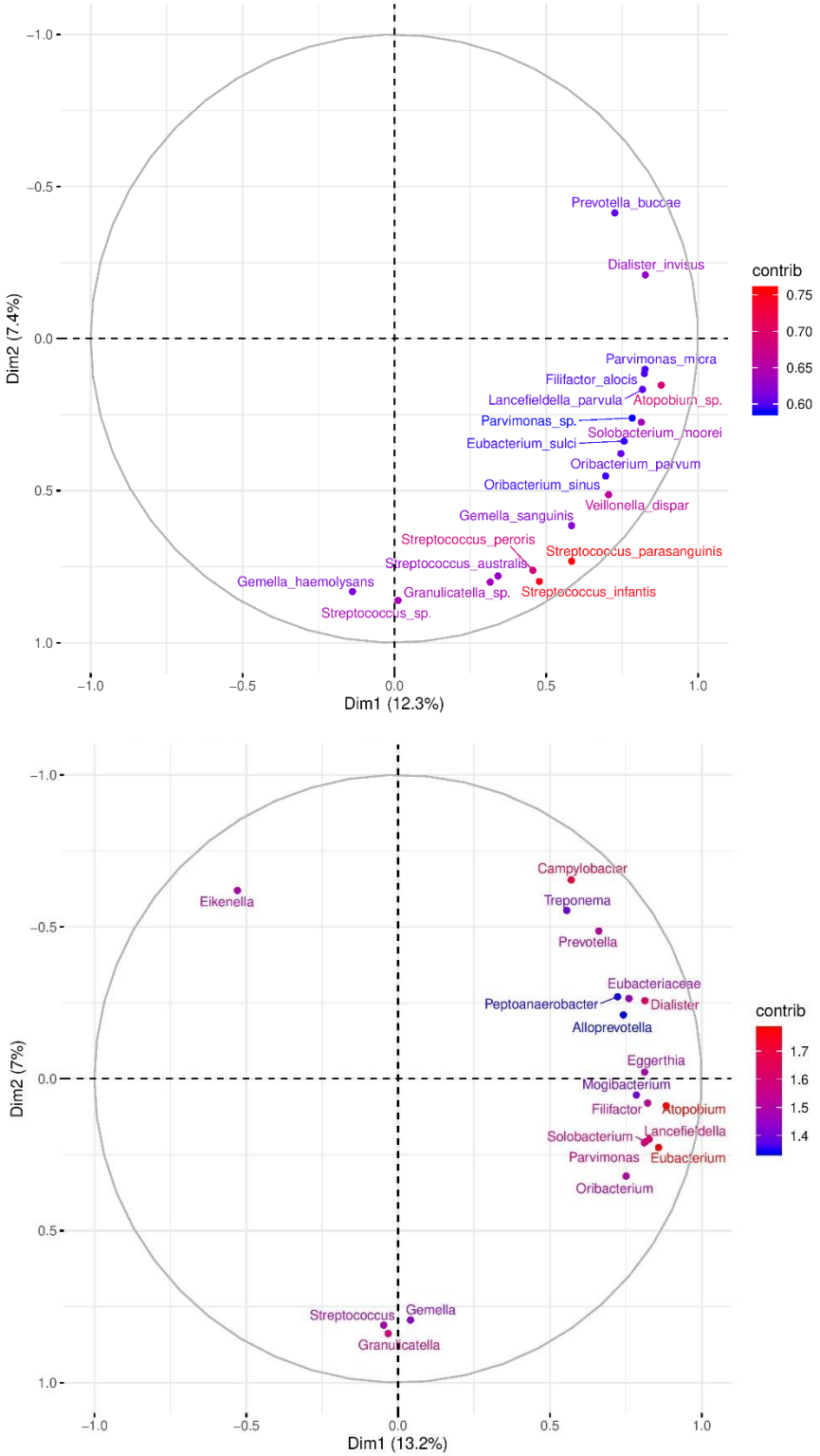


Figure S7 | Loading plots (positioning of the variables species and genera) of the principal component analysis based on microbial species (A) and microbial genera (B) data from 70 Southern African individuals. For both PCA the 20 species/genera with the greatest contribution to PC1-PC2 are represented. Species are coloured by contribution as indicated by the legend.

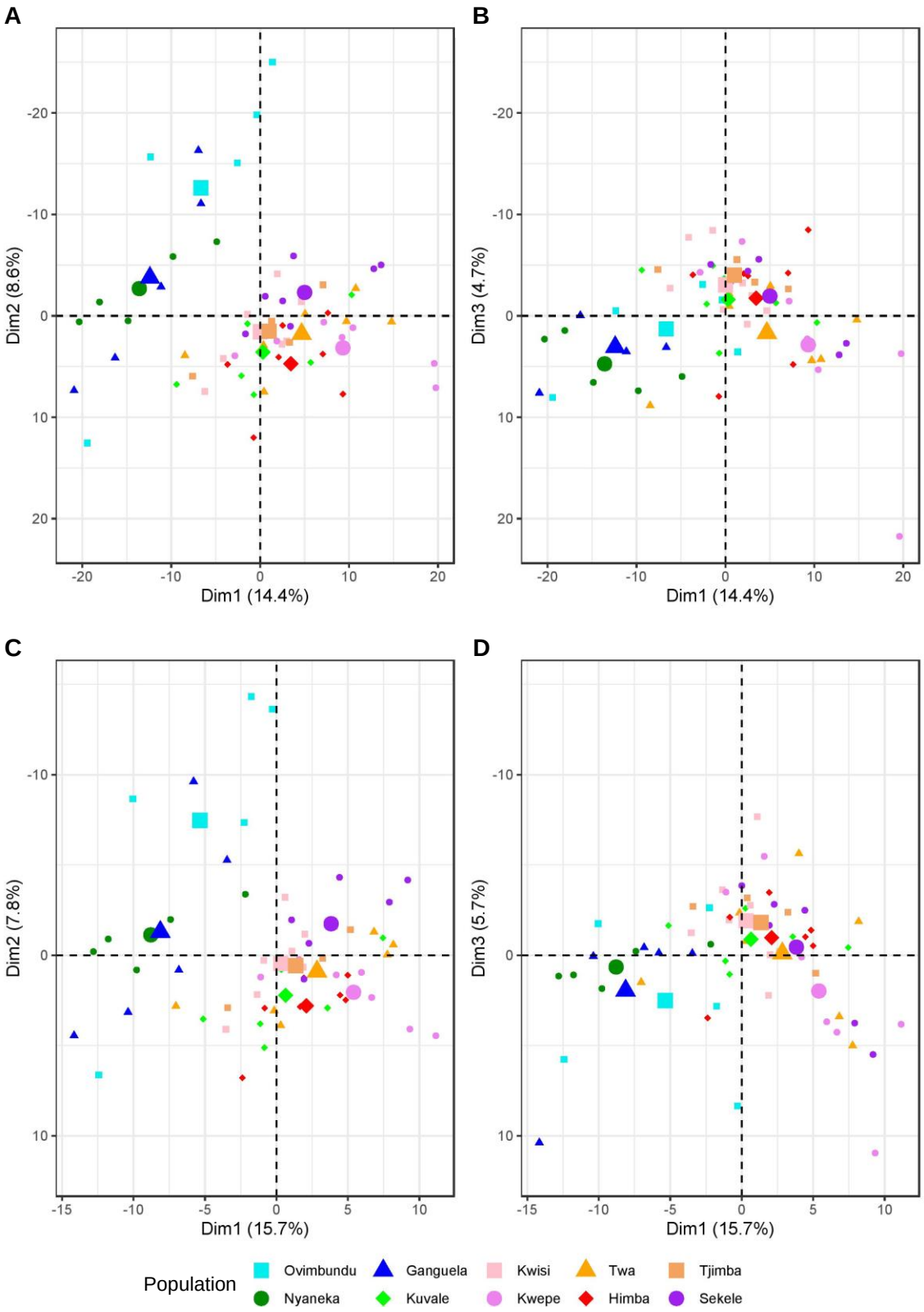


Figure S8 | Principal component analysis based on microbial species (A and B) and genera (C and D) data from 62 Angolan individuals. A and C represent PC1 and PC2 whereas B and D represent PC1 and PC3. Individuals are represented by a specific combination of symbol and colour representative of the population. Population midpoints are indicated with larger symbols.

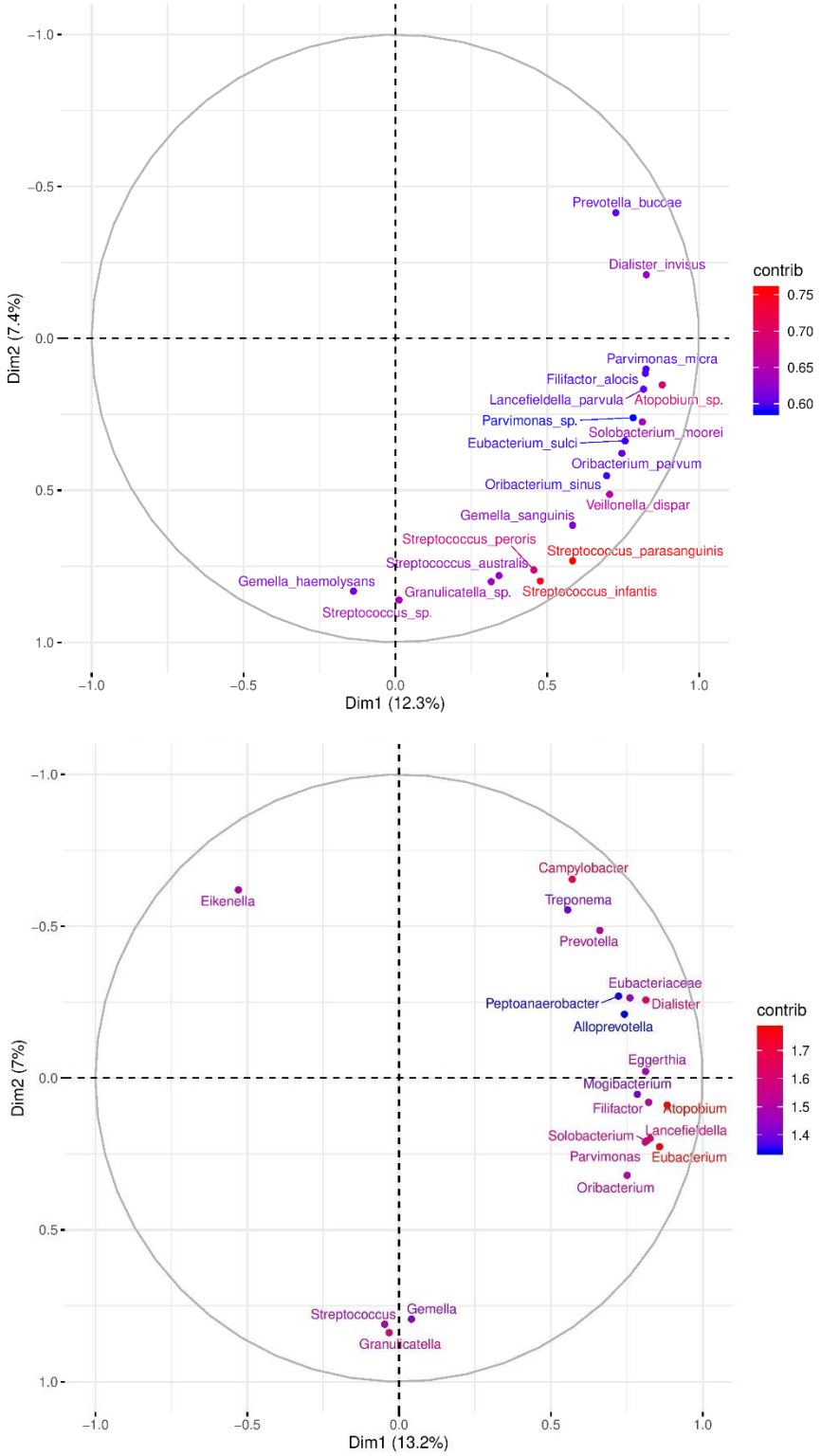


Figure S9 | Loading plots (positioning of the variables species and genera) of the principal component analysis based on microbial species (A) and microbial genera (B) data from 62 Angolan individuals. For both PCA the 20 species/genera with the greatest contribution to PC1-PC2 are represented. Species are coloured by contribution as indicated by the legend.

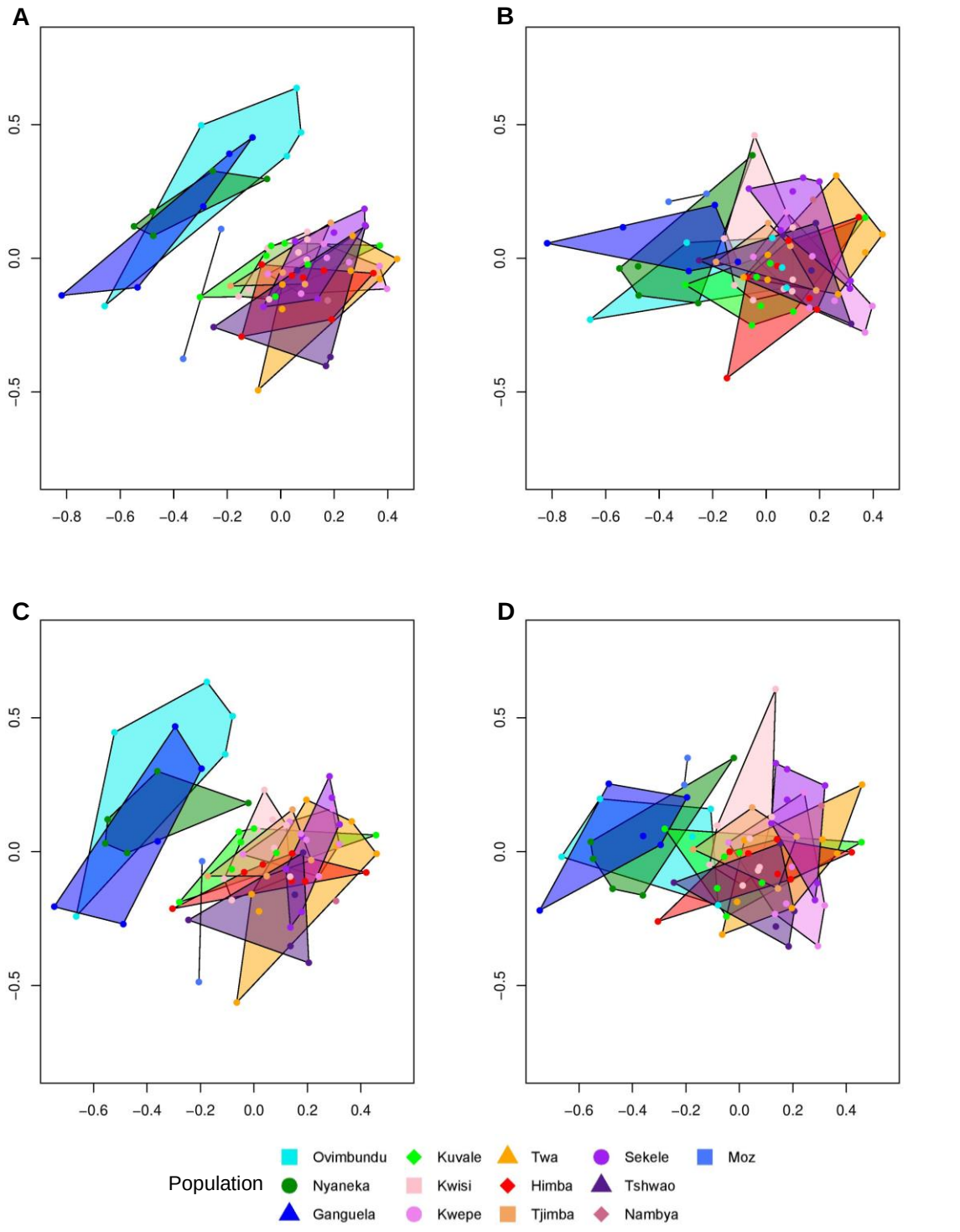


Figure S10 | Non-metric multidimensional scaling based on microbial species (A and B) and genera (C and D) data from 70 Southern African individuals (in both cases a three-dimensional NMDS was built). A and C represent dimensions 1 and 2 of the NMDS, while B and D represent dimensions 1 and 3. On the species based NMDS (A and B) the stress level was 0.128 whereas on the genera based NMDS (C and D) the stress level was 0.144. In order to better understand the positioning of the populations in the low-dimension space, the individuals with more extreme positions in each population where connected by lines forming population polygons represented by a specific colour as indicated by the legend.

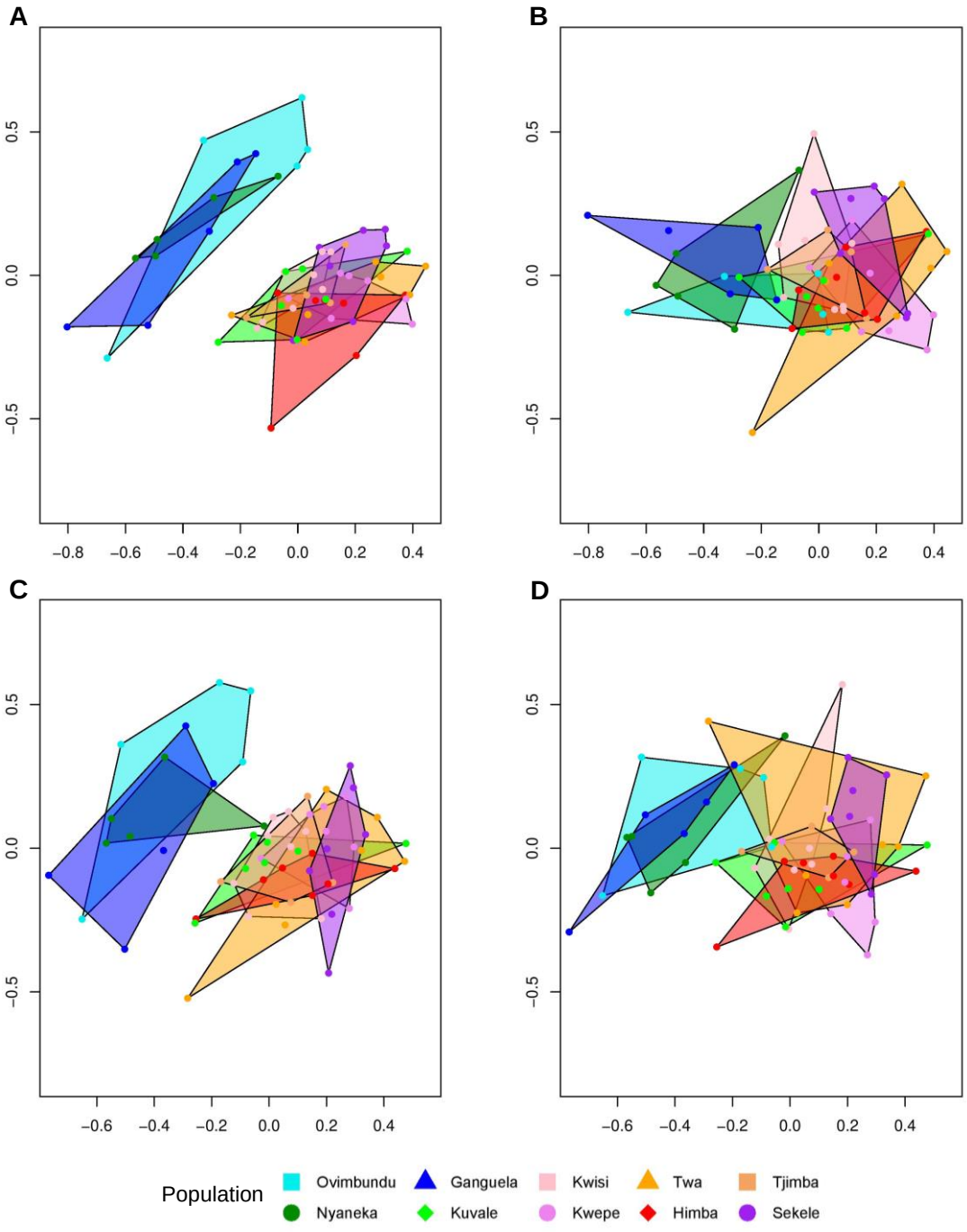


Figure S11 | Non-metric multidimensional scaling based on microbial species (A and B) and microbial genera (C and D) data from 62 Angolan individuals (in both cases a three-dimensional NMDS was built). A and C represent dimensions 1 and 2 of the NMDS, while B and D represent dimensions 1 and 3. On the species based NMDS (A and B) the stress level was 0.12 whereas on the genera based NMDS (C and D) the stress level was 0.134. In order to better understand the positioning of the populations in the low-dimension space, the individuals with more extreme positions in each population where connected by lines forming population polygons represented by a specific colour as indicated by the legend.

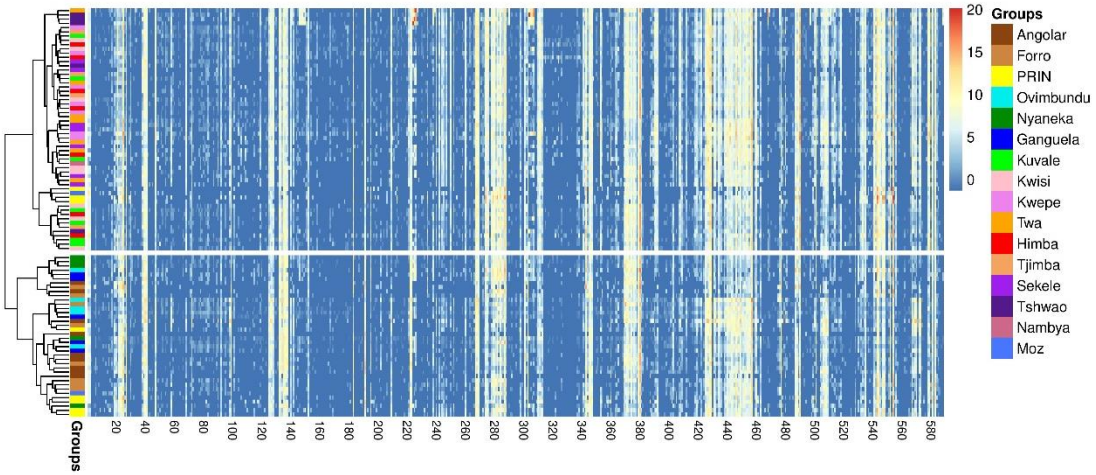


Figure S12 | Heatmap of species abundance in 95 African individuals. Individuals (rows) are coloured in relation to their population.

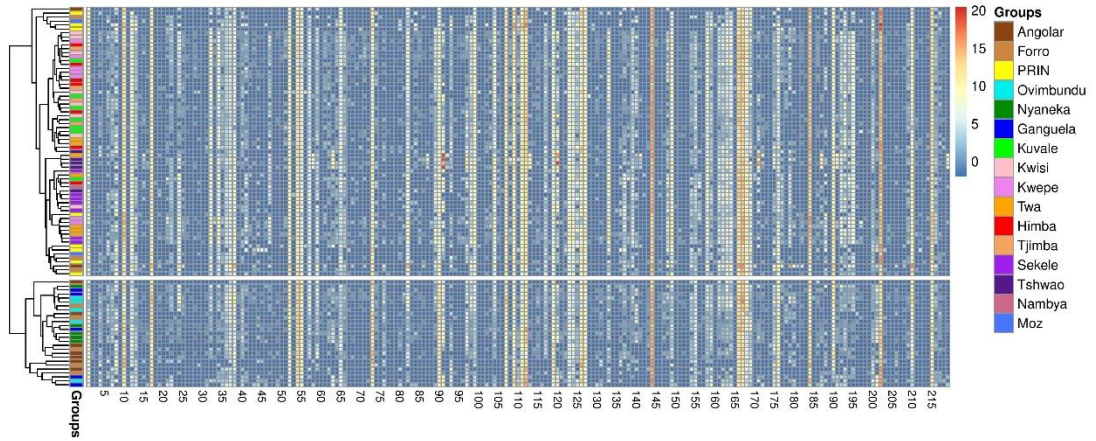


Figure S13 | Heatmap of genera abundance in 95 African individuals. Individuals (rows) are coloured in relation to their population.

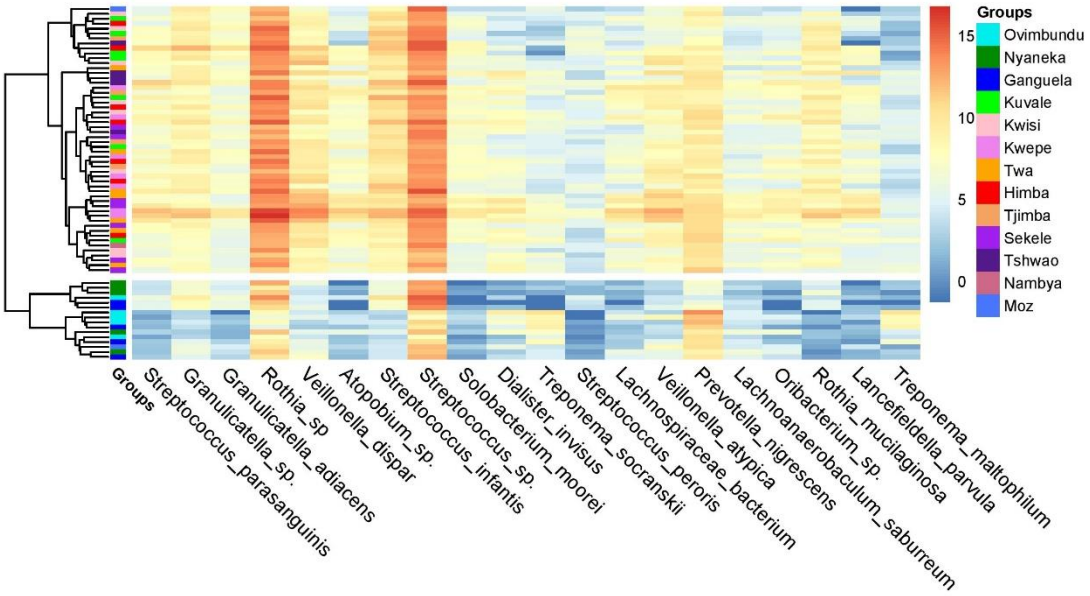


Figure S14 | Heatmap of species abundance in 70 Southern African individuals. Only the 20 species with the greatest contribution to principal component 1 and 2 of the respective PCA are shown. Individuals (rows) are coloured in relation to their population.

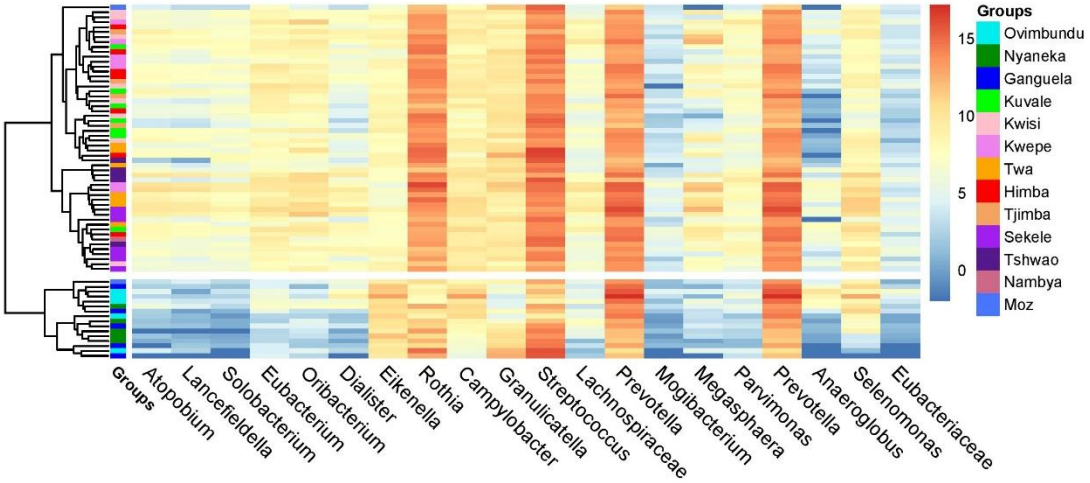


Figure S15 | Heatmap of genera abundance in 70 Southern African individuals. Only the 20 genera with the greatest contribution to principal component 1 and 2 of the respective PCA are shown. Individuals (rows) are coloured in relation to their population.

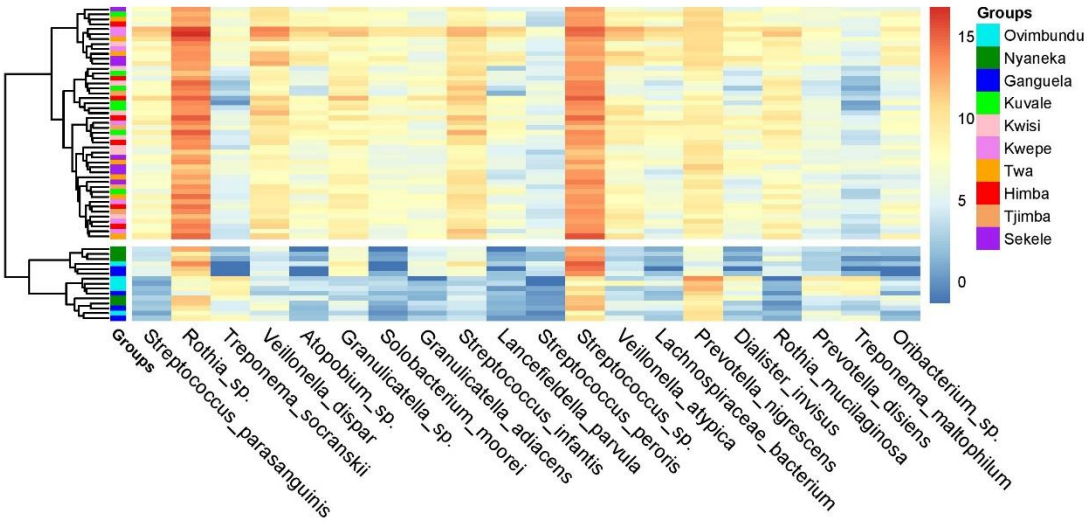


Figure S16 | Heatmap of species abundance in 62 Angolan individuals. Only the 20 species with the greatest contribution to principal component 1 and 2 of the respective PCA are shown. Individuals (rows) are coloured in relation to their population.

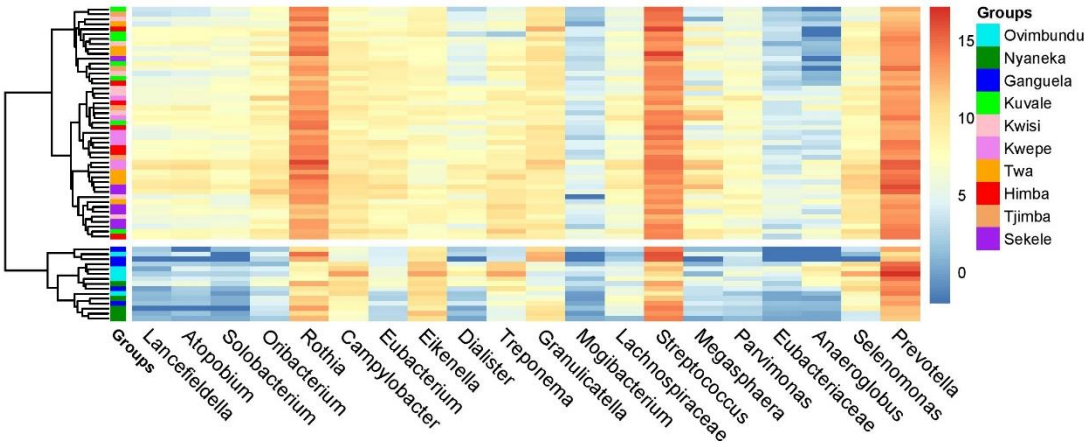


Figure S17 | Heatmap of genera abundance in 62 Angolan individuals. Only the 20 genera with the greatest contribution to principal component 1 and 2 of the respective PCA are shown. Individuals (rows) are coloured in relation to their population.